PROCEEDINGS OF A SYMPOSIUM ON THE USE OF EVALUATION BY FEDERAL AGENCIES

Edited by
ELEANOR CHELIMSKY

April 1977

THE METREK DIVISION
OF
THE MITRE CORPORATION

THE NATIONAL INSTITUTE
OF LAW ENFORCEMENT AND
CRIMINAL JUSTICE

WELCOME AND INTRODUCTORY REMARKS

HERBERT D. BENINGTON, Vice President and
General Manager, METREK Division of
The MITRE Corporation

On behalf of the METREK Division of The MITRE Corporation, I welcome all of you to this three-day symposium on "The Use of Evaluation by Federal AGencies." I particularly want to welcome Jerry Caplan, Director of the National Institute of Law Enforcement and Criminal Justice, LEAA (whom I will introduce shortly) and to thank him and the Institute for joining with us in sponsoring this meeting.

The number of people who have shown an interest in this meeting is but one piece of evidence illustrating what a timely, lively and important subject this is. I think the gross reasons are obvious. During the past decades, we have seen a phenomenal increase at all levels of Government in the amount of support for major social programs. In some cases, the need for these programs has been driven by past neglect and deficit. In some cases, they have been driven by a strong desire for higher standards of service and higher public expectations. In some cases, both of these factors have been driving forces. However, the important observation is that there has been a dynamic increase in the numbers of social programs implemented by all levels of government over the past ten to fifteen years.

More recently, I think even the most optimistic have begun to agree that there are questions as to how effective these programs are. Some of the issues concern quantitative questions of efficiency, of productivity, of waste. Perhaps more importantly, qualitative questions have arisen as to whether some of these programs are even achieving the

the most important objectives for which they were established or whether, in fact, some of the programs are counterproductive. Thus, we have seen a growing emphasis on the need for program evaluation throughout the whole system. This concern is evident among those who develop and manage the programs, in the watchdog offices within the Government, on the Hill, and in some of the special and public interest groups--all of whom are looking over our shoulder. All of these various activities are represented at this meeting. We thank you for coming to exchange your experiences and your aspirations and look forward to gaining new perspectives from your discussions.

The MITRE Corporation grew up dealing with formal complex systems. We really started back at the Massachusetts Institute of Technology (where the MIT in MITRE and the MT of METREK come from) during World War II, working with the radiation laborabory. The challenge at that time was integrating radars, communications, very simple computing devices, men, and other machines into systems that would meet some objective. As we were conducting this work, we came upon a most important conceptual realization: the absolutely critical role that feedback plays in a system. As a matter of fact, one could easily say that the theory of systems is the theory of feedback.

By way of illustration, one of my recent loves is molecular biology, an area where enormous progress has been made in the last couple of decades. As we start to understand the simplest bacterium as it is just sitting there at rest, we discover that there are thousands of feedback loops within that system. Life itself is a process of feedback. Certainly in our society we have all sorts of feedback mechanisms, having recently seen one used by the public to send a new President to Washington.

Given this perspective, the significant questions that this symposium will address are: How does one deliberately and formally, explicitly and publicly, design feedback into all key elements of a major social program? How does one include feedback between the different elements of that program: in planning it, in developing it, in managing it, in implementing it, and in operating it? What will the feedback mechanisms be?

The use of evaluation as a management tool was formally initiated by the Department of Defense and the National Aeronautics and Space Administration. The incorporation of evaluation into DOD and NASA programs was relatively simple--these programs dealt with high technology but both the developer and user were within the Federal Government. A program can be readily identified as "successful" when an astronaut comes back safely from the moon having accomplished the program objectives within 10 percent of the original budget.

The programs that we will be discussing during the next three days are vastly more complex. These programs consist not only of technological components, but involve new developments in management techniques, in dealing with the market, in international implications, and in the creation of new social institutions. As I said, it is not only the Federal Government, but all levels of Government which are involved in these programs. The programs may have different objectives, targeting segments of society ranging from the affluent to the ghetto. It is within this context of multiple levels of decision-making, differing roles and participants, and the numerous and frequently divergent objectives which change over time, that evaluation must be designed and conducted.

Another important point is that evaluation cannot be undertaken after the fact. It is not adequate to set up the Assistant Secretary for Program Evaluation at a late date, give him some very bright people, and let them go at it. Evaluation is a component that must be built into a program from the very beginning.

With those few words of perspective, let me now introduce our key speaker for this morning, Jerry Caplan, who is the Director of the National Institute of Law Enforcement and Criminal Justice at LEAA, the organization which joins us in sponsoring this meeting. I have here a long resume for this young man. I hope it is not a resume occasioned by this time of the evolution of Washington. It points out that he graduated from Northwestern University with several degrees, most recently his law degree. He studied political science at Yale. He has received too many awards to enumerate. He is a member of too many professional associations to list, and he has too many publications for me to even count.

Before he came to the Institute, he was a Professor of Law at Arizona State University. Previously, he was General Counsel to the Metropolitan Police Department in Washington, which must have been an interesting job. He was also General Counsel of LEAA. He has a very bright and distinguished career. Jerry, we thank you for coming and for sponsoring this symposium with us.

THEME OF THE SYMPOSIUM

GERALD M. CAPLAN, Director,
National Institute of Law Enforcement
and Criminal Justice

Thank you for the generous introduction. I want to welcome all of you here. As this conference has developed, I have been impressed with the high degree of interest expressed by individuals who wanted to know more about it and to join us. From the beginning, the conference seemed a very good idea: an opportunity for those of us engaged in evaluation to get to know each other and compare notes. I expect that all of us will learn something. At the very least, we can commiserate with each other about the problems we are enduring.

I don't yet have a good sense of the mood of the people here. For those that I work with on a daily basis at the Department of Justice, LEAA, and the Institute--my sense is that a great deal of goodwill exists on the part of those who sponsor research, those who are responsible for implementing its findings, and those in the research community that actually do the work. This is an important aspect of the current climate in this emerging enterprise.

At the same time, I sense an undercurrent of disappointment that is much more difficult to articulate--a sort of vague feeling that we may not be quite on center track. Perhaps as the conference develops, we will be able to articulate it more precisely and see whether it is a misperception on my part or whether it reflects some deeper concerns that should be faced.

My mood is mixed. For our forthcoming discussions, it may be worthwhile to stress the negative side. I am apprehensive about what I perceive as an evaluation "boom." That may be very much a

Justice Department perspective. It emerges from the legislation, from the media, from criticism, from our own internal efforts and from the views of certain individuals within LEAA. I sense evaluation today as something that inherently smacks of virtue. It's the "right thing to do," an article of faith. It is easier to pretend to abide by it than to dissect it and try to analyze its strengths and weaknesses.

We are talking less about evaluation as a way of illuminating problems, putting them in sharper focus and plainer view, which is the way I personally like to think about evaluation and research; and more about finding "answers" and "solutions," which I consider to be, at the very least, overblown rhetoric, and, maybe something more disturbing. It may be a kind of optimism that to me would be a harbinger of things to worry about. At one level, it could signal abdication of managerial responsibility, a way of relieving people who should be in charge of responsibility and turning it over to somebody else. More important, ballooned expectations often carry with them an aftermath, a kind of hangover, which may bring important evaluation efforts to an unwarranted early termination or reduction in scope, funding and/or enthusiasm. I believe we have a responsibility to nurture the process in a very responsible way and not make too much of the child. I am less concerned about making too little of what we are bringing into the world.

Rather than attempting an overview of government-wide use of evaluation findings, I'd prefer to give you the LEAA perspective, and we can see later on to what extent that is typical or not.

Within LEAA, particularly since 1973, we have consciously, conspicuously, and earnestly turned more and more to evaluation as

a way of acquiring information to improve programs, to improve policymaking, to set priorities, and to plan our future research agenda. While it may be too early to judge, it's nonetheless fair to say now that evaluation has seldom furnished us with the kind of knowledge and information anticipated by LEAA decision-makers. I stress the word "anticipated," because I think you could argue that we got our money's worth, that we received pretty good work for the dollars invested. However, program expectations (or statements of expectations by program developers) were often exaggerated and then set against the built-in caution of the evaluation methodologies; it is that chemistry, I think, which compounds the impact. Whether or not our expectations were too great and/or our evaluations too timorous, we have seldom gotten the knowledge we anticipated.

Furthermore, when knowledge has been forthcoming, we have not often used it; and I'd say we haven't used it for very good reasons. These reasons don't have to do with caprice or whimsy or individual idiosyncracies but rather with the way bureaucracies work: the problem is that evaluation studies which take some time tend to get out of sync with the natural flow and needs of the agency. This is why utilization is a major problem for us, and I think we are not alone in this but are typical.

Despite this less than optimum experience, however, the pressures from Congress continue to mount to do more evaluation. This is not surprising since the Congress appears to use the word "evaluation" quite differently than we. Very different images come to mind with that word, "evaluation." It is not a problem in definition. It is simply that Congress talks about it one way, and we haven't quite got the hang of what they really mean. I think the Congress uses evaluation in a more casual sense of making disciplined judgments

about what you are about, how you are spending your money. More and more, we refer to experimental design, process, outcome and sophisticated statistical techniques for analysis. The Congressional demand for evaluation seems to me to reflect in turn a general public demand for agency accountability, a way of finding out what we're up to. This public demand has been translated into a political requirement for more relevant, continuous and effective Congressional oversight. This requirement and the demand for evaluation also stem from increased budgetary pressures. One reaches out for evaluation as a way of making more sense of the allocation of Federal resources or the relative merits of competing priorities.

In this context, I think it made sense to MITRE and the National Institute to invite you to join with us in looking at these kinds of problems: the difficulties that those of us who represent agencies have experienced in acquiring needed technical information; the kinds of strategies you as researchers have evolved in trying to meet our needs; the perspectives that we have all developed in dealing with each other and what we can do about them; how, in fact, we perceive each other--as friends and colleagues, as people in an alliance, or in an alliance of enemies, since an adversary notion is inherent in all this. There's no doubt that some of this is difficult, perhaps impossible to fully unveil. My own instinct tells me that the more we open it to scrutiny, the less dangerous things get, and the more likely it is that our relations will smooth out over time.

The hardest part for us here, I think, will be to develop some sort of prognosis for what we can expect (if I can steal a phrase of Elly Chelimsky's[1]) in the way of progress under pressure.

---

[1]Head of Program Evaluation, The MITRE Corporation, METREK Division.

Three workshops have been scheduled. The first will focus on the relations between agencies and researchers. Working Panel II will look at the actual problems of evaluating program effectiveness and offer a summary of the present state-of-the-art. Working Panel III will focus on the utilization problem--that is, what do we do with evaluation after we've got it? Based on an agency experience, this group will recommend methods for enhancing the process or, at least, understanding better the process by which evaluation findings are funneled into policymaking or are not funneled. I hope that we can lay out the current expectations of evaluation by the different audiences involved; the experience in using evaluation results; what it is realistic to expect in the future and produce a set of recommendations.

We recognize the arduousness of the tasks. They deal with basic concerns: our agency's well-being, our company's well-being, our university's survival, expansion, termination, own own personal stake in these things, the way we are valued, job security, promotion. They also deal with other kinds of problems, such as maintaining scientific rigor that is often a nuisance in the real world, or adapting the state-of-the-art in social science research to the kinds of tensions I referred to earlier that are inherent in program and agency politics.

There has been a great deal of recent experience, so I believe now is just the right time to get together and talk about these issues. Evaluation offers a real potential for illuminating major program questions about how we function, for addressing the trade-offs among conflicting priorities, for more rational and articulate policymaking, and for improved agency performance. I look forward to participating and contributing and being instructed during the next three days.

MR. BENINGTON:

Jerry has always been working in areas which, from my point of view, are obviously very important and generate great enthusiasm in most people. I have always been delighted with the extent to which he is prudent, cautious, and thoughtful. I thank him for the low-beat start.

Microphones are available to anyone who wants to make a challenging or critical or hopeful comment on what Jerry has said or to ask him a question on what he has outlined as going to happen. Please approach one of these microphones, state your name and organization and let go.

PARTICIPANT:

Cork Grandy, MITRE/METREK[2]. I am wondering, Jerry, if you could expand a little bit on why you think evaluation results have not been more frequently or more fully used. The question of getting out of sync with other agency processes caught my attention.

MR. CAPLAN:

Let me step back for a minute. When I first had the chance to begin to make remarks such as these, a senior official at the Department of Justice called a number of us together and said it is always good to begin your opening remarks anywhere with some self-deprecating comments. All of us at the Department of Justice--this was a few years back--thought that was excellent advice; but try as we might, we could never come up with such anecdotes. The point was that this would make the audience more sympathetic to what you are saying and less likely to criticize. Despite the very

[2] Vice President, The MITRE Corporation, METREK Division.

able people associated with us there, there was a paucity of imagination; and that may explain that this group doesn't need to be wooed by a self-deprecating story, but is naturally generous in asking tough questions. I want to encourage you to ask them.

My answer to your tough question is a cop-out and perhaps appropriately so at this time. What I described may be unique to the Deaprtment of Justice or LEAA. Perhaps other people really don't have these kinds of problems.

Let me spell out a couple of them. I think that we have an especially tough job compared to other agencies. Criminal justice evaluation is much more difficult for several reasons. One is that it starts off with the law itself. The law is vague in its demands, or has multiple demands with built-in tensions. We want to have laws and we want to arrest people for breaking them, but sometimes we don't want to arrest people. We want to make statements, grand statements, about how we see ourselves. We have many laws on the books that nobody ever takes seriously, and we would be profoundly shocked if they became part of the criminal justice process. The law itself is made up of an ambiguous set of dictates. Stemming from that, I think there are inherent tensions within the criminal justice system that make a lot of sense; but they are very difficult to articulate, and we shy away from it. For example, arrest is often a rational act from the point of view of a police agency, but prosecution would be irrational from the point of view of a prosecutor. Prosecution is often just the right thing to do from the point of view of the values of the subsystem of criminal justice, but conviction would not make sense. The same thing is true with conviction and incarceration. There are discontinuities that inhere in the system. To what extent they ought to be is not my point, but rather the fact

that the tensions and discontinuities are there. I don't see other systems quite as much at odds with themselves about where they want to go. That would be part of the problem.

In a much more narrow sense, the LEAA statute itself is brewing with problems of just this kind. We are supposed to be innovative, which means that once we hit some winners, instead of further developing their anti-crime potential, we have to move immediately to something new and creative. This is not merely a tendency to keep moving so nobody can hit you. This is what we are supposed to do. It's true that we have some mechanisms for translating the winners or the good news that has come from our studies into other institutional channels. But there is a tension between developing long-term strategies and continually being innovative. So that it often happens that by the time an evaluation comes into being--a solid, fine evaluation--the same people may not be there. If they are, their interests may be different. Legislation may have changed. The state of crime may have changed. The world may have changed in terms of interest in that program. That's what I mean by being out of sync.

MR. BENINGTON:

I was remiss in my opening remarks in not thanking Elly Chelimsky, who I am sure some of you have met and talked to on the phone and who has been the spearhead of this conference within our organization and, I think Jerry would agree, on the part of both of our organizations.

MR. CAPLAN:

Yes, of course, I said it first. I should say in appreciation, along with Herb, that the controversial parts of my remarks were suggested to me by Elly.

MR. BENINGTON:

Any more comments or questions? All right, then, Charles Grandy, my cohort here at METREK, will make some overview remarks.

SYMPOSIUM ORGANIZATION AND OVERVIEW

CHARLES C. GRANDY, Vice President,
METREK Division of The MITRE Corporation

I would like to add my welcome to those of our previous speakers. One of my chores is to mention to you a couple of administrative items that may be helpful to you during our conference. You may have noticed when you arrived that we do have arrangements in the anteroom and in the entrance to the building for telephones and secretarial assistance for those of you who may be expecting messages or need other help in running your other businesses while you are giving your time and attention to our proceedings.

We also have transportation to the luncheons and the dinners which are at locations other than our building. While there are parking facilities at the hotels and restaurants, they are sometimes congested and limited. Many of you, I know, are acquainted with the area and may prefer to take your own wheels. But if you choose not to do so, we will have shuttle buses getting back and forth to the hotels and to the restaurants.

One other comment. We are making recordings of our proceedings, either through a tape or stenotypist, and it may be well for you to be aware of that. We will prepare a transcript of this material, and we will of course check things with our speakers before we go into print.

You have seen a number of METREK staff members here who have the yellow or orange symposium badges; they will be happy to help you with any questions or problems that you may have during the conference. They are identified that way so that you can easily pick them out.

Our agenda, I think, is familiar to you. Your presence here is an indication of your interest in the material. We have a couple of changes in our program which I will mention shortly. The response to this symposium has exceeded our fondest expectations. We set out to have a modest group of about a hundred people, and the response has been in excess of 600. Today we are some 200 strong. I encourage and in fact exhort you not to let this size, particularly in our Working Panels, be an inhibition to earthy and free-wheeling discussion and commentary. The purpose of the conference will best be served if we can be calm, cool and professional, but enthusiastic and vigorous as well.

We have a good variety of speakers and topics on our program, and we have a very diverse attendance, with individuals from the Executive Departments and Cabinet agencies, from GAO[3], the Congressional Budget Office, OMB[4], from the Congress and the staff of the Congress, and we have researchers in the area of evaluation from many walks of life and many professional interests.

I think the image that each of us has about the future importance and uses of evaluation needs to be brought out. The special problems that the users, the representatives from the agencies and the program managers in those agencies have had with evaluative research need to be aired fully and adequately and to be considered and understood by researchers. The real point of our symposium is to try to stimulate some improvements in existing linkages between the decision-makers

[3]The General Accounting Office.

[4]The Office of Management and Budget.

in the agencies and the evaluators. The ultimate objective of this improvement ought to be better, more effective and more useful programs in our society.

So evaluation is a tool, and we very much hope that out of this conference we can get a better understanding of how that tool can be better used on both sides of the equation by the users and by the evaluators.

Our approach to the symposium as you know is to have, first on our program, the presentation of agency experiences, views and needs. We have in this morning's program, Chaired by William Carey, nine representatives from agencies covering health, energy, crime, education, what have you. These folks have been asked to tell us about the experiences that their agency has had, both good and bad: the successes, the shortcomings, approaches and strategies--but most importantly, the needs that they have for information pertinent to their decisions and their plans.

We have tried to stimulate and augment the total contribution of this part of our program by conducting--in advance of the symposium--interviews with each of these speakers. We think this is a somewhat unique feature for a conference of this kind. Those interviews were recorded and transcribed, and copies have been distributed to the members of the Working Panels.[5] We think this will make the focus of those panels more specific at the same time that it provides a more complete background to the views and needs of these agencies than we could possibly get in the fifteen-minute discussion scheduled for presentation today.

---

[5]These interviews are at Appendix II of this Volume.

I noticed in our agenda that we have billed Mr. Carey as Executive Director of the American Academy for the Advancement of Science instead of the American Association for the Advancement of Science. One of our participants commented to me earlier that this was not an entirely inappropriate error since sometimes the AAAS contributes in the same ways that the Academy of Engineering and Academy of Science do. It is nevertheless an error.

Among this morning's people, there is a replacement for Robert Knisely, who was unable to be here. His replacement, Mr. Tom Kelley, is a program analyst in the Office of Program Evaluation for the Department of Commerce. We are especially pleased that he can pinch hit, since he was a major participant in the interview with the Department of Commerce to which I referred earlier.

After the presentations of agency perspectives, which will establish one background for us, we will move to a presentation of researchers' perspectives. This will include six presentations by ten investigators, again in a wide variety of areas, as you can see from the agenda. The introduction to this part of the program will be given by Dr. James Abert. This will provide baseline information from the research viewpoint so that we can then move into the real guts of the program, which we expect to be the Working Panels.

These Working Panels (and indeed the whole symposium, as Mr. Benington mentioned), were conceived and organized by Eleanor Chelimsky. Jerry Caplan has outlined their charges. Panel I, improving the user/evaluator interface, will be Chaired by Clifford Graves. Working Panel II, on improving the definition of evaluation criteria, is Chaired by Marcia Guttentag, and the third one, on improving the utilization of findings, will be Chaired by Blair Ewing.

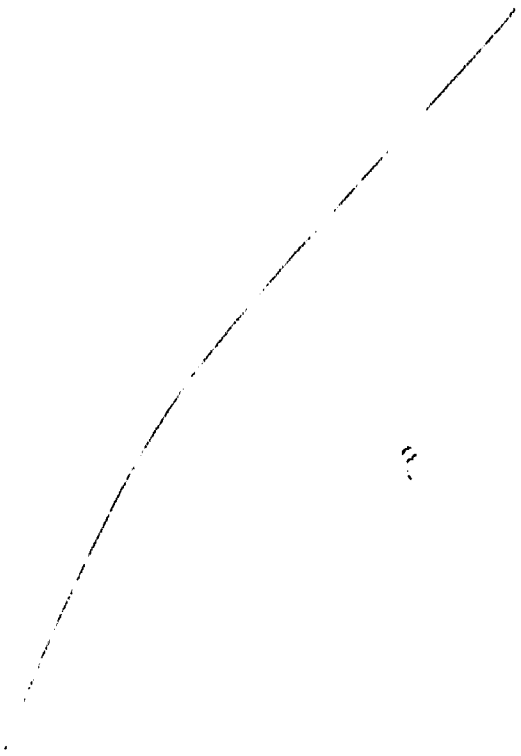
Our attempts to get something useful and productive out of the conference may have led us to give you the appearance of an overly structured program in the Working Panels. We have delineated some issues with which the panel chairmen are familiar and which we hope can be a useful starting point and focus for these panels. We do not, however, want this to be a rigid session. It ought to be flexible and imaginative, a broad-band, free-wheeling exchange of information. The Working Panels will provide reports for us on Friday.

Many of you, in registering, indicated your interest in participating in one or the other of the panels, but I think not all of you have done so. We have 55 people indicating an interest in Panel I, 67 in Panel II, and 45 people in Panel III. Those of you who may not have signed up can either let us know your selection, or follow your interest to the panel location of your choice when we reach that point in our program. Everyone is welcome at these working panels, and our intention and hope is that you all will be able to take an active part in them, despite their large size.

We want to prepare from this conference a report that is going to have some real impact in the real world. We have an ambition that I think is timely. As Jerry Caplan pointed out, it's a hot topic and one where I think fruitful progress can be made, even if under some pressure. Our report will present, not only a transcript of these proceedings, but also an analysis that will try to compare and contrast the approach to program evaluation taken by various agency decision-makers, to delineate the methods that have proven useful in the past, and to make recommendations for the future.

I think if we can do these things, we will be able to realize our ambitions of a symposium that is something more on the Washington scene than just another conference. We at METREK and at the National Institute are truly pleased to have so distinguished a group of

speakers, agency representatives, luncheon and dinner speakers, researchers, members of working panels and vigorous participants in this conference. From your enthusiasm, your knowledge and your work, I think, will spring a highly successful meeting.



INTRODUCTION TO THE AGENCY PERSPECTIVES PANEL

WILLIAM D. CAREY, Exeuctive Director,
The American Association for the Advancement of Science

MR. GRANDY:

As we now continue with our conference, it's my pleasure to introduce to you Mr. William Carey who will be the Moderator for the presentations of agency experience and perspectives. Mr. Carey is the Executive Director of the American Association for the Advancement of Science, a position which he has taken after six years with the Arthur D. Little Company, and before that, a very long and distinguished career in public service. He is a native of New York City, holds a number of degrees from Columbia and Harvard University and has served in a variety of public posts in the Bureau of the Budget, a number of White House task forces and Cabinet Committees and is very widely experienced in the subject matter at hand. It's a real pleasure for us to have him here to moderate this part of our program.

MR. CAREY:

Thank you, Charles. What a line-up we have up here this morning!

I think we are here to look the facts in the eye. I think that, as has been said here, evaluation is one tool that can be helpful in assuring the quality of governments and administration; and as a state of mind, I think that is fine. I am not sure that it's enough, and I don't think we have come here to hold a self-congratulatory feast about the whole business. I thought that some of the remarks I heard when I came here this morning were very, very direct. I have also had a chance to read the results of the interviews and as I read them I was absolutely fascinated. They are superb. They are clear. They are candid, they are honest, they are great. There is a lot of truth. But I sense there is a great deal of confusion of terms and meanings. Evaluation, analysis, social science research, accountability, control,

policy research, management--all these terms criss-cross the traffic of evaluation. I think they mean different things in different contexts and at different levels of management and different levels of decision-making. There are some contexts in which evaluation seems to thrive, and there are others where it withers. Some places, it is a way of exercising power. In others, evaluation couldn't matter less.

So its uses are debatable, and they are varied; and the track record is a mixed one. I think all these descriptors fit, but they don't fit uniformly. That is part of the problem. We can look at the very short history of evaluation in public management, and I was there when it began; and we can be dazzled if we choose to be by the appearance of a very pretentious industry which has come into being because a market was created for it. On the other hand, we can look and we can see something else. We can see a very encouraging development in the direction of a new kind of public management which is exciting, which is still having growing pains, but which is pretty darn sure to make it in the end.

I tend to put my own value on it for what it's worth. It seems to me that in spite of the failures and the frustrations and the scarcity of conspicuous successes, what we have here is one of those very rare examples, certainly in my experience, of long-range investment; and it's the kind of thing that I would say in time will be ranked with the emergence of the Executive Budget 55 years ago, and with the beginning of macroeconomic policy some 30 years ago. It could be, and I tend to believe it will be, the third leg in the array.

So we are going to look today at evaluation and its credibility, where it has come thus far from a standing start, what the expectations are, whether they are realistic or inflated, whether there's too much propaganda behind them. We are going to try out what we have learned. We want to know whether evaluation has made a difference; and if so, what is the quality of that difference and what is the prognosis.

The way to do this, I think, is to talk to the people who are doing it the hard way. We are going to start this morning--as I say, we have nine speakers. The last thing I'm about to do is to parade them before you one after another. What I would like to do instead is to run them at you, perhaps two at a time, and then take a few minutes to get questions because if we wait until the ninth is finished, it's going to be pretty hard to catch up with number one. I want to get number one to work pretty hard. So that is the way we will play it. I hope that the speakers will do their best and at the same time try to contain their remarks within a 15-minute time-band each.

First we are going to hear from Sam Seeman from the Department of Health, Education and Welfare, speaking on the evaluation of health care delivery. Sam and I used to see more of each other when he was the Executive Director of the National Capitol Health and Welfare Council of which I was at one time a Board Member. It's nice to see him again. Sam.

THE AGENCY PERSPECTIVES PANEL

I. HEALTH CARE DELIVERY

ISADORE SEEMAN, Acting Deputy Assistant Secretary for Planning and Evaluation/Health, Department of Health, Education and Welfare

It's great to be here if only to see Bill again after a long time. I am the most fortunate one in this room. I am number one of the nine. I am reminded of the convocation of ministers where the first one who presented the invocation for it said, "We have a great conference here, and the Lord bless the first speaker. Give him a silver tongue so that his words come forth and bring us light. And Lord bless the second speaker and give him wisdom so that the message is very clear. Lord bless the third speaker, inspire him so that his message sends us forth on our way; and Lord have mercy on the last speaker."

I looked at just the first seven speakers in the morning, and I was struck by the fact that the first speaker is from HEW and the seventh, John Evans[6], is also from HEW. I am not flattered, however, because if this is a sandwich, you and I are the bread, John, and the rest of them are the meat. So I guess we have a challenge.

MR. EVANS:

Especially since it's high quality meat.

MR. CAREY:

Yes, but it's peanut butter in between.

MR. SEEMAN:

I hope the record will show that.

---

[6] Agency Panel Member John W. Evans, Assistant Commissioner for Planning, Office of Education--Department of Health, Education and Welfare.

What is the best way that I can spend a few minutes with you and share some thoughts about health evaluation? I could take 15 minutes and tell you about the structure of health evaluation in HEW. I could tell you how many dollars we spend on it. I could tell you the number of evaluation studies that are in our library. I will tell you all of those things if you ask, but I think it would really be pretty dull to hear that recitation from Health and EPA and LEAA and so on.

It seems to me what I might best do is to spend a few minutes on the major perspectives of a guy who is working personally and professionally in evaluation. Then we can get into more specific questions if you wish.

We do know, as has been said already this morning, that we are dealing with social institutions which are relatively new in the history of this country. I don't think that there was very careful evaluation when the Pilgrims landed here as to how the trip went and what the expectations were. In fact if there had been, they might well have turned around and gone back.

But that is 200 years ago and more. We have been investing in social programs in a significant way for only a very short time and assessing how they are doing is still a younger activity. Therefore, my thoughts are framed in terms of frontiers for evaluation. I would like to suggest four frontiers that trouble me, frontiers to which I give a lot of thought as I try to do the job of providing some guidance to the evaluation of health programs in HEW.

The first is the need for us to understand and define the necessary balance among those activities that make a difference in program management and policymaking and legislative development. We should recognize that evaluation is a part of these processes, but only a part. I think there has been (and this was suggested this morning but I would underscore it) too much of a tendency to put evaluation on a pedestal. It is a great thing. It is a marvelous thing. I am sure each of you at one point or another has had the kind of query we've had coming, for example, from the Appropriations Committee: "Will you please send us a list of the evaluation studies that led to the termination of particular programs?"

We haven't sent them any list because there aren't any such studies. Evaluation studies represent one way of getting information, but there are a lot of other ways of getting information, too. Unless we appreciate that, we can go off the deep end. So, evaluation is one ingredient in a stew, or it's one weapon in our arsenal, or it's one tool in the whole tool kit. I think we need to recognize that, relate it effectively to other tools, and appreciate that it has a place but that it's not necessarily the final answer.

The second frontier, and a more important one in my view, is the frontier of professional leadership. It seems to me, in spite of what I say about evaluation being only one of the tools, that it ought nonetheless be a more important tool than it is. It's tough to make it so, and it's tough to make it so because to do the kind of evaluations that ought to be done and, when you have them, to make something happen because of them, takes guts. It takes leadership. I think we need to give more attention to that quality.

It's not easy to ask the most pertinent questions about a program. In fact, it's tough to get the most pertinent questions, first of all, listed on a piece of paper, so as to make some judgments about what are the most relevant issues, what are the key questions. What happens much more frequently, certainly in our experience, is that a lot of peripheral questions will get asked, but as for the gut questions, it's tough to get them listed. When you do get them listed, then it's tough to get somebody who has some responsibility for the program to agree that he'll let you in the door to take a look at those tough questions. Then it's not easy to get your own staff, or an outside organization under contract, to face those tough questions, to address them, much less answer them.

What I see involved in this issue of the kind of leadership that is needed to press forward with a more effective use of evaluation--and here the first question of balance comes into it again--is the fact that what we are faced with is a world in which political decisions and political forces play a very strong role in the outcomes of the programs we are concerned with. I am unhappy when these political factors are the only ingredients in the decision-making process. It has to be balanced with the rational element. This is tough to do, and that is where the leadership and the guts come in. It seems to me that we will have, and ought to continue to have, political factors affecting decisions as to whether this piece of legislation goes through or not, whether this program gets a lot more money or is modified in a significant way. That ought to be one of the ingredients. The other ingredient ought to be objective, analytic, rational, knowledge-based information; and evaluation can help produce that information. But it won't if we just coast along and say, "Congress will do what it wants to do anyway, so it doesn't make much difference."

The third frontier, as I see it, is a frontier of incentives, incentives to do good evaluations and then to do something about them. This seems to be an awfully important area that we have not developed sufficiently. What are the rewards for good evaluation, and what are the punishments for poor program operation? They are all too few. What will you do when you have the facts? Take welfare, for example, we have plenty of evidence, and you have heard it thousands of times that the present welfare system leaves something to be desired. It's a mess. You have heard about the welfare mess for how long a time? It's documented. We are still trying to get some welfare reform. We don't have a successful experience in getting welfare reform. What are the incentives to make the changes? Unless we analyze those, we are not going to get the kinds of changes we want.

Take the case, for example, of the Community Mental Health Center program that HEW has had in operation for about a decade. Suppose we did a very clear analysis and evaluation of that program. Let's say we found it cost five times as much money--and this is hypothetical--five times as much money for an encounter between the professional and the client in a Community Mental Health Center as it did in private practice. Or let's say you find that clients come into the Center and they drift off, they don't really stay long enough for any real effectiveness. What would we do about it? Do you think Congress is going to turn that program off?

As a matter of fact, what really happened was that, instead of saying the Community Mental Health Center movement is ineffective we should terminate it, we tried the approach of saying, "The Community Mental Health Center movement has been very effective and therefore we ought to terminate it as a Federal program. We have proven that it's good. Therefore the States ought to pick it up and finance it, and we need no more Federal money in it." That didn't work either.

It seems to me that, in fact, the agency in the Federal structure that has the greatest potential for improving the usefulness of evaluation is the Civil Service Commission. If they could only devise the kinds of rewards and punishments for effective program management and for sound policy decision-making, I think we'd get better program management; and we'd get sounder policies carried out.

The fourth frontier is in the methodological area. It seems to me that we often hide too much behind the fact that we don't have all the refined methodologies to do the evaluation studies that we want to do. I would try to strip away that shield because, in many cases, I think it's not a valid one. But in one fundamental way, it is. I'd like to see more work done on this frontier.

Let's say you really want to determine the influence of a particular program that you are sponsoring: for example, a new piece of legislation for health planning across the country. You want to know what difference health planning makes in expenditures for health care in the country which are skyrocketing to the point that everybody is terribly worried about it. What you want to know is, what difference did the Health Systems Agencies that are now being created across the country make. You don't want to know what difference it made that there was a rate-setting agency side by side with the HSAs. You don't want to know what difference it made that a particular physician has his own orientation, and that it's his work and not the work of the Health Systems Agency that brought down the length of stay in the hospital, for example. We need to be able to control for these things if we want the answer to our question. We don't presently have adequate tools in the social fields to do that kind of a controlled study. I think we may never get the ability to do controlled studies of that kind. In a laboratory you can say, we will give the placebo to this

group, and we will give the experimental drug to another group. You can do that on a small scale, whereas with large social activities you obviously can't. But we can and should, I think, find proxies for the controlled effort. Otherwise, we don't really know what difference the particular activity we are measuring made and what the external forces were.

Well, those are some thoughts about evaluation from one who is trying to work in this area. Fortunately, we are in a little better situation than that of the rooky who was just learning parachuting and asked the instructor, "Sir, how many parachute jumps do you have to make successfully in order to win the insignia?" The instructor replied, "All of them."

MR. CAREY:

Thank you, Sam.

II. ENERGY RESEARCH AND DEVELOPMENT

J. FREFERICK WEINHOLD, Director,
Office of Evaluation,
Energy Research and Development Administration

MR. CAREY:

We are going to go right on now to Fred Weinhold. Fred is the Director of the Office of Evaluation in ERDA.

MR. WEINHOLD:

Good morning, and thank you. I feel a little bit like the virgin coming in to lecture a group of parents on the joys of parenthood. All of you people have worked in evaluation for a while, whereas I have had the title, Director of the Office of Evaluation, since July 1st. They haven't given me any staff yet, and it was not until last week that I finally got my action plan for doing things approved. But we have got an Office of Evaluation now.

Let me give you a little perspective on ERDA and on our evaluation problem, which I think is somewhat different from that of most of the rest of you. I am not going to say we have a harder problem or an easier problem. I have no idea about that, but there are a few technical and physical differences that do make it a different problem.

First of all, ERDA is a new agency, as you know; energy research and development within the Government is a relatively new function. Five, six, seven years ago when I started working on this, the total Federal budget in the whole area was some three or four hundred million dollars a year. There were three or four programs: the breeder

reactor, fusion and a couple of little coal projects were what we had in the late 1960s. This has mushroomed into a $2½ billion-a-year program. We have got lots of new programs in it, in conservation and geothermal and solar energy. The main thrust of the agency and of our work has not been on evaluating, but on starting new research and development programs. The thrust of our work over the last two years has been focused more on planning and the analysis that goes into planning than it has on post facto evaluation or evaluative research of any kind.

When you look at what it would mean to evaluate some of our research and development programs particularly, you find there is not that clear a distinction between planning, analysis and evaluation. Theoretically we are trying to develop new options for the country or develop new technologies which would provide insurance. Most of these won't have any impact in the energy economy for 15 years at the near side--the big programs, 20-25 years. Some of the other programs are 40 years away. The obvious way of evaluating, of going through the program, doing it and seeing what the results were is totally impractical here because the results of the evaluation 40 years from now would be meaningless.

So the question (and our most serious problem) is trying to figure out ways of doing meaningful evaluation and looking at the programs--all in a prospective or future sense. What have you done in the past year and a half or two years or four years, perhaps, that will make a difference in 30 years that is different from what the situation would have been if you hadn't done it 30 years in advance? This gets you into a lot of analysis questions that are not that different from the analysis that goes into starting up the program.

One is immediately struck with the need to use modeling tools, future projections and other similar instruments in trying to estimate system performance that would work into the future. You do this same sort of thing when you are buying into a program that you do three or four or five years later when you are trying to evaluate it. The work we have done follows closely along these lines.

We are also planning to look at how individual programs have been managed over the three or four or five-year period to achieve these goals and look at comparisons of various alternative R&D strategies. Do you proceed with five different technologies that lead to the same market or the same thing in parallel, or do you end at a certain level of expenditure? Do you focus all your funds in one particular area? These are the critical issues which need to be decided in energy research and development.

One specific example. During our budget process last August, we were looking at long-term energy options. That is where the big money is, in fusion, solar electricity and the breeder reactor. If you sit down and look at how many systems are being pursued in research and development in this area, the number is somewhere on the order of 14 or 15. In the breeder area, you just predominantly have one; but in fusion, there is magnetic fusion and laser fusion. In each one of these, there are three or four different approaches, hopefully aimed at the same target.

So the evaluation challenge we have there lies in trying to sort out when is the right time to cut down on the options. How many do you try to keep open knowing that none of them will really prove out, one way or another, for 25 to 40 years?

Those are some of the issues that we have. Up to now, I have been focusing on the energy research and development and demonstration part of our program. That is the name of the agency and what people think about, but it is not our only activity by any means. We have three or four other major activities within the agency that have posed some special evaluation problems.

First is national security in the area of weapon research and development, testing and production. We are the ones who figure out what new warheads should be built. It's a chicken-and-egg situation with the Department of Defense--who decides on the requirements and what can be done; but we do the research and development, the testing and building of the nuclear warheads.

Another area that is fairly big is basic work in biomedical and environmental research, as well as in the physical (energy-type) research area. How does one evaluate research? I don't know whether we will come to any discussions of that sort of thing this afternoon at the conference or not. I'd be interested to get any feedback that people here may have in that area.

Then the one that is really a tough one to evaluate is high energy physics. How good is it learning about some of these black holes and what does such knowledge really do for the country, and how do you evaluate progress in that? I think we'll sort of hang back and wait a little while before we get into that sort of thing.

Then the final area is really a business. We are in the business of enriching uranium. It's a fairly big business--about a billion dollars a year. One could conceivably evaluate this with business criteria, profit and loss and things like that. However, there are political and other difficulties in doing that in the agency, particularly when we

are trying to sell the country on privatization, or on allowing new industrial firms to compete with us. Congress has been rather reluctant to provide the incentives to industry to let industry compete with us in our business here. So that is a different sort of problem.

Those are the types of issues that we are facing which are somewhat different from the social research, or more people-oriented types of evaluation I think most of you are struggling with. But perhaps a comparison back and forth between some of the issues will help me and perhaps help you in this discussion.

What have we been doing up to date? We have been doing a number of analytical pieces, part of the planning effort, that also feed into the evaluations. We call them market studies and macroscenario work. We try to project into the future without new technologies, and then project into the future with new technologies and see how these different futures stack up and try to get some estimate of the value of the work we are doing and try to do some sort of cost-benefit work.

Cost-benefit analyses are probably not too bad for the technologies that would have some impact in the 10 to 20, maybe 25-year time frame. When you use OMB's 10 percent real discount rate, you find that it says the present generation is not that interested in saving your grandchildren 10 or 20 percent on their electric bill. That is what the 10 percent discount rate says. I guess that is probably true. The only reason we are interested in doing that sort of work is to protect our grandchildren against cataclysms caused by not having energy systems available. This gets you into an insurance-type problem and risk. It's a little bit different than the normal cost-benefit work.

We have just started in-depth evaluation, that is, we have laid out program evaluation plans and sent them to OMB a week ago. Our approach is to run evaluations essentially in-house, with small teams augmented by consultants and contractors, to try to look at individual programs over a three to six-month period asking the questions, "Does the program make sense if it succeeds?" "How well is it being managed internally?," taking into account the funds that go into it and the risk involved. Is it a good risk versus the cost decisions being made in it? We plan on running and developing detailed plans in advance of these and getting them approved by the Administrator and the program people that are involved in this before starting it. We end up with written reports and take a four to six-month period for evaluation, then look at the buy-in decisions. Does it make sense to escalate this program from a modest research stage to the development stage or to the demonstration stage? We hope to schedule and tie these evaluations in to major decision points in the programs. In energy research and development, there are some clear steps. They vary almost by an order of magnitude in the funding that goes into them. You work at the bench level for millions of dollars. At the prototype level for tens of millions. Demonstration plants in a lot of these are at the hundreds of millions. When you start talking about commercial plants in nuclear power, in coal gasification and some of the other biggies, you are talking a billion dollars a shot. There are some clear economic decision points, and we hope to tailor the evaluation schedule so they would feed into these.

Our approach is to feed the results into the program, to the assistant administrators and to be working with the Administrator on this.

One of the issues that has come up and that we are struggling with now is, how does program evaluation--as I have laid it out--fit into the overall energy analysis plan or the agency analysis plan? What has occurred each year is that everybody starts looking at the total agency program and says, "Gosh, I'm worried about fusion this year. There are lots of questions and issues with it." The question that comes to us then is, of all the bag of tricks we have (from developing program plans and strategies to doing special studies, from bringing in outside review groups to conducting program evaluations), which do you apply to this particular program or concern in a particular year keeping in mind that, if you tried to apply all of your tools to a particular program, nothing would get done and nothing very clear would come out. So we have had to be selective and we've tried to focus on a particular time in the year when we would decide which tool gets applied to which program, and try to make some decisions on that. We are in the process of that now, in fact.

To wrap up quickly, then, we are tending to focus on relatively large programs from our central view, looking at things at the hundred million dollar a year level, rather than applying these sorts of techniques to small programs costing $10 million or less. We don't think it is appropriate to put a couple of man-years of effort into evaluating something which doesn't support a sufficient payoff from evaluation to make the amount spent on that evaluation worthwhile.

We also see that there are needs within the agency for what I would call project evaluation and audit. There are other functions that go on within the agency, not at the staff level, but within the programs that do continue on. That is the part of the agency's overall effort.

So these are our plans and hopes right now. But they are surely up for grabs as the new administration comes in and transition takes place. I don't know five months from now what I would say to a group like this. I thank you.

III. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

MR. CAREY:

I said that after each brace of speakers I'd take a couple of questions from the floor. We have heard from Sam Seeman. We have heard from Fred Weinhold. Are there questions from the floor at this point before they get away from us?

PARTICIPANT:

Yes, Jim Robinson with the Department of Labor. First a quick observation, then a quick question. It seems somewhat distressing when you are looking especially at the delivery of social services to see a panel that is so unrepresentative, first of all, of population and secondly, of the clientele to whom a lot of these services are delivered.

The second thing is, could the panel direct itself to the basic question of whether Government evaluation seems to be overly input-oriented? Probably this is because we have our best fix on inputs. Yet, it is whatever comes out of the pipeline and how it really impacts people that really determines whether or not the Government is performing the basic function for which the taxpayers are paying. What I am particularly thinking of in HEW is, for example, now that the courts have turned down the Hyde Amendment on Medicaid abortion, should the Department realistically go forward with an appeal?

Secondly, on vaccines--does it make any sense to shovel vaccines out into the delivery system when we are now getting feedback which shows that large percentages of the population, especially in the black community and certain other communities, are just not

interested in taking these things because basically they don't trust the Government and they don't trust the quality of the vaccines? In other words, should we stop looking at the inputs? It doesn't matter whether or not the vaccine performs effectively in a laboratory if the people aren't going to come and take the vaccine. In other words, it's the ultimate output rather than the input which counts. How do you get evaluators really thinking about that, especially when the evaluators very seldom are representative of the population? Most of them have forgotten what a lot of problems are like at the very level where the services are being received.

MR. CAREY:

I am going to score that as four questions. Sam, lots of luck.

MR. SEEMAN:

Let me start with the last one. You catch me on a reasonably good plane on this one. The way we got into the swine flu vaccination program, which certainly is something that is on many people's minds today, is something I won't describe now. The Secretary of HEW (who at one point had to decide to approve, or not approve moving ahead with it and decided to move ahead) is seriously concerned about how that decision was made and what the results will be. Two days ago, I finished Version Four of the draft memorandum on a comprehensive evaluation of the swine flu program. I think and hope there will be such an evaluation. Certainly we want to see that it occurs--those of us in the Office of the Secretary who have some responsibility to advise the Secretary, and the Secretary himself, want to see such an evaluation.

It will not be only an evaluation of the laboratory aspects (that is, was the vaccine safe and effective) but also, the delivery aspects.

Some of us have been quite seriously concerned for a couple of years about what appears to be a significant decline in attention to vaccines of all kinds by young parents. Immunization is essentially a childhood need. All of us can get flu vaccinations, but diphtheria and pertussis and tetanus are for kids. Parents aren't doing it nearly as much as they used to. Polio vaccination is not what it ought to be. We have been concerned about that.

Again, you are into the blend of objective and political factors. On the one hand, we ask objectively, "What data do we need? How do we get those data?" and then the political forces come into play. When I say political, this isn't Democratic and Republican. There are political forces in every agency. The political forces in HEW said, "We hear you. There seems to be some problem, but there are bigger problems to worry about"--I am oversimplifying--"Let's not pay too much attention to the decline in immunization."

We had a conference this weekend that said, "Hey, it's more of a problem than we think." I feel there will be a blend of political factors and objective factors that will lead us to do something more, about immunization in this county.

MR. CAREY:

In HEW, is evaluation concentrated at the front end of the Department, or does it go like the streaks in a marble cake all through the Department and particularly out in those regions where people are and where impacts are delivered and felt? What about the organizational extremities of HEW, your regional offices?

MR. SEEMAN:

The evaluation effort is exceedingly splintered--as splintered as the programs are. The major HEW component dealing with health is the Public Health Service which itself has six agencies in it; then there is Social Security which runs Medicare and the Social and Rehabilitation Service which runs Medicaid. There are seven major units. Then each of the public health service agencies has bureaus and divisions. I would guess there are no less than 20 evaluation offices in HEW centrally. Then there are 10 regions, and each region has an Associate Regional Director for Planning and Evaluation. But they have very small staffs. Thus, evaluation is done throughout the Department.

I think that is a strength and a weakness. On balance, I'd say it's probably more of a weakness than a strength. It diffuses the effort. It doesn't give you enough of a component at any one level to really tackle the effort as seriously as you'd like. A staff of two or three people can't do the kind of work that it takes to deal with immunization and the whole Medicare program. We are scattered and splintered

MR. CAREY:

Do you think that Bill Morrill--I guess he is Assistant Secretary for Evaluation over there--would give consideration to pushing more of the responsibility for evaluative work into the regions?

MR. SEEMAN:

I'm not so sure I would push that much for the regions, Bill. The regions have a role. Regions want to get more into it, but the regional offices of HEW are the place where the rubber meets the road, where the programs get implemented. What the regions want to do is study national policy. I don't think that's the most appropriate place

to study policy. Our policies are essentially national policies. Whether we ought to have a certain type of home care demonstration effort in the State of Washington is no different than whether we ought to have it in Maine. But whether the way it is working can be improved or not is another question. A regional office could do something about how a program is working.

MR. CAREY:

I guess the reason I asked is that there is a gentleman from Georgia going to be taking charge of affairs around these parts pretty soon, and he made quite a point in his campaign about decentralization and getting things out into the grass roots and so forth. It would seem to me that it might be thinkable that a regional strategy of evaluation, not only of operations but of what ought to be done and how it ought to be done, might very well further that goal.

Is there another question?

MR. EVANS:

While he is going to the microphone, let me just interject for a second and add a quick comment, and suggest that that topic you just raised is an important one for this conference to consider in more detail. I think Sam has given a reply that I personally would be inclined to agree with. The general assumption that the anti-government theme of the Carter campaign (and indeed the anti-government, anti-centralization feeling generally) should lead one to think of evaluation as a function for decentralization should, I think, be examined and questioned because one of the problems in the decision about where evaluation is located is the important issue of objectivity in evaluation, and the extent to which an organization

or program officials should evaluate themselves. This is one of the dangers that is likely to occur in the kind of system which becomes totally decentralized. There are a lot of pluses and minuses on both sides of that issue, and that, as I say, ought to be one for additional discussion.

MR. CAREY:

That's all right if you assume objectivity is a function of geography.

MR. EVANS:

I would argue that it's a function of program responsibility or involvement.

COMMENT FROM FLOOR:

It needs to be evaluated.

PARTICIPANT:

My question springs from Mr. Weinhold's comments. I'm Tom Richardson with the Department of Commerce. My overall impression from your comments, which also lead me to try to generalize from that to what we are doing in evaluation, tends to go as follows. It seemed to me that your discussion of the various ERDA programs tended to follow the go-go syndrome. It appears as if the name of the game in ERDA is to crank out all these systems and possibilities that will generate energy in the years to come; that the thrust is to see how well we are doing and how quickly we can do it and how we can reduce the cost--that kind of thing.

Obviously, there are certain negative factors tied up in the energy field. It seems to me that to include those in the evaluation would tend to reduce the attractiveness, and hence also the

support for the agency's thrust. I guess my question goes to this: Shouldn't evaluation be more squared away? Shouldn't it deal with the bad as well as the good and not be totally supportive of whatever the agency's mission is and hence a kind of bureaucratic enhancement, a sort of strengthening of that particular agency's push? Is my point clear?

MR. WEINHOLD:

Yes, it's a good point. I think the first step we need to take when we go to evaluate any program is to say, "Okay, assume that the technologists or proponents are successful in meeting their goals and targets--is the program still one that the country would like or should have or makes any sense?" I think that's the place I want to start in each one of these, saying we assume success in the way people have laid it out. Does it still make any sense for the government or the country to try it? Are the environmental or the economics or the other attributes of it useful or not?

I guess the overall energy growth demand question is a very complicated one. I don't think we or anybody could attempt to say, "Okay, it's going to be good for the nation to grow at 4 percent in energy growth per year or not good, as opposed to 2 percent." What we have tried to do in our market studies and in some other efforts is to say, "Okay, look at a couple of futures. A high-growth future and a low-growth future. Does this technology make any sense in a high-growth future? Does it make any sense in a low-growth future? How does this technology look vis-a-vis the others if you have a nuclear moratorium or if you don't have a nuclear moratorium?" I think those are the ways we try to raise these questions. When you start looking at the breeder and fusion and questions like that, you are trying to look at various ways the

nation could go in 20 or 30 or 40 years and try to see how the technologies stack up in it. That is what we are trying to do. It's a pretty fuzzy subject, though, to try to come down with anything that is meaningful. The ranges are so wide.

MR. CAREY:

Thanks Fred. Let's go on. We are coming to a subject, crime prevention and control, that we all know how to deal with. We are all very competent evaluators. We have all the answers. Anyway, let's listen to Dick Linster, who is the Director of the Office of Evaluation at the National Institute of Law Enforcement and Criminal Justice, within LEAA.

IV. CRIME PREVENTION AND CONTROL

RICHARD L. LINSTER, Director,
Office of Evaluation, National Institute of Law Enforcement and Criminal Justice,
Law Enforcement Assistance Administration

Bill's last remark was just slightly inaccurate. There is at least one person in this room who doesn't know all the answers in criminal justice evaluation, and that's me. It was about three years ago that the then Attorney General, who was very interested in evaluation, spoke on it to LEAA officials and said that evaluation means he wants to find out what works. As one of the technical middlemen who was supposed to operationalize that concept, it made me very nervous. I am still very nervous about it.

I'd like to describe a little bit of what I think are the basic problems that LEAA faces in carrying out an evaluation program that makes some sense and that, in the spirit of this conference, leads somewhere in decision-making. First of all, we are a block grant program. The bulk of the money is allocated to the states by formula. This isn't just a cosmetic arrangement. It was very much based on a philosophical spirit when the agency was created: Congress didn't want the Federal Government telling the states and cities how to go about controlling crime. Clearly the question of whether or not we would be gradually moving towards a Federal crime control system, a Federal police force, was one of the things that was a real concern in the debate over the LEAA legislation. It was very clear in the way the agency was set up, with the state planning agencies being effectively independent of the Federal LEAA. Congress seems somewhat more ambivalent about this now. I think

the explicit demand for evaluation in the 1973 and 1976 reauthorizations is some indication of that ambivalence. But the fact is that that is the way the thing was set up, and Congressional demands for greater programmatic accountability can create federal-state tensions if they are regarded as an encroachment on states' decision-making autonomy in this program. In particular, then, it can be extremely difficult to get information about what the block grant money is doing once it goes through to the states.

The reasons why it is so hard to make clear, succinct and scientifically defensible statements about what general effects the LEAA program is having are not, however, entirely "political." Evaluation of and within LEAA is also faced with very fundamental technical and conceptual problems.

First of all, LEAA money is roughly a 5 percent add-on to the money that is already being spent on the problem of crime and the operations of criminal justice. Grantees are diffused all over, not only geographically all over the United States, but all over the criminal justice system. And not just the formal criminal justice system. Citizen groups are also included--citizens particularly interested in doing something about the crime problem in their local communities. So that the substance of what is going on under LEAA grants is just as diffused. Then also we are talking about a lot of grants that go out in the $10,000 or $20,000 range. There are relatively few grants, when you consider the LEAA program as a whole, very few grants that go out in terms of three or four hundred thousand dollars, that is, larger individual single grants.

That is one problem. But possibly a more major problem is that of simply conceptualizing what it is that we would like LEAA to be doing, no matter how it's structured, as a block grant program or as a set of categorical grant problems. For purposes of evaluation, global statements of agency goals must evidently be translated into an adequate system of observable measures of change and that can be far from trivial--even within the context of a particular program area.

For example, we are now working on design of an Administrator's discretionary grant in the area of court delay. That seems like a very simple sort of thing to evaluate. You can presumably go in and measure some statistic reflecting what the time of trial is now. Then, when some type of program has been undertaken in a court under an LEAA grant, you can go in and measure that time of trial later. If there has been a reduction, you say the program has been a success. But obviously, the existence of a delay problem is only a symptom of some larger problem in the system. One can evidently clear the dockets if they are overcrowded by all sorts of measures--dismissals, plea bargaining. But those measures may not correspond very well with what the whole system of criminal justice was intended to do.

LEAA started out, I think, with a clear understanding that the goal of the agency was crime control. We had to bring street crime down. We had to bring it down through provision of Federal assistance. But the defined goals of the agency have changed somewhat in the time I have been there. The formal goal--this was originally in the Act--the formal goal, the emphasis in what is presently being stated about

the LEAA program, is now pretty much "system improvement." But "system improvement" itself requires definition. Obviously, what is meant is that the system, after you have done something to it, is a better system than the one you had before. It's improved and presumably you have some concept in mind of what you mean by improvement.

I don't think we can disguise the fact that people still think a criminal justice system ought to do something about crime. The criminal justice system is in essence the formal mechanism by which our society tries to keep crime at some optimum level.

Still, one can talk about "improvement" in other senses. One can talk about improvement in the sense of efficiency--essentially maintaining a constant level of effectiveness but at a reduced cost. States are going broke, they say. Cities are going broke. A police chief in a major city has to get his budget justified, get money to pay for patrolmen and pay for new equipment. He may want to expand his program. The question then, a question of efficiency, can be clearly a goal of the LEAA program and, in consequence, this is a proper theme for evaluation. But it's very similar to crime control in the sense that we really don't know very much about how to measure efficiency either.

Here I think the problem is that we really don't understand the dynamics of the criminal justice system as a system. This sometimes is described as a non-system, but I tend to think that that is probably inaccurate. Sub-system goals may appear to conflict but that may mean only that there is a hierarchy of goals.

What I am thinking about is what Jerry Caplan was touching on earlier. That is that the apprehension and prosecution goals of the police and of the prosecutor are quite distinct and quite different from the justice goals of the court system. I still think that overriding all of this, however, is the idea that criminal justice is established in the United States or in any country to provide a mechanism for crime control in the society.

Well, in terms of evaluation, the conceptualization of the system, if we had such a thing, would be a distinct blessing. We would be able to say, for example, that we understand the dynamics of the system so that when a program is put into operation in a court, we can talk about what the implications are in terms of changes in the plea bargaining rates, changes in the incarceration rates, what the impacts are going to be on the correctional system, on the parts of the process of criminal justice that takes the offender from time of arrest to time of release from the system.

We have some descriptive models of this, of course. Models that are empirically based, that are essentially linear flow models that have taken a criminal justice system in a given jurisdiction and have collected the data that measures branching ratio. Where are the branches in the system, if you try to follow the offender through?

What we really need is a much better understanding of the whole dynamics of the criminal justice system so that we have some kind of a basis for limiting an evaluation, for saying that an evaluation of this program doesn't really have to look for secondary effects all the way down from the stream and all the way upstream. It can

simply look at a particular intervention. That is one of our major problems. The conceptualization, the theory of system dynamics is not terribly well developed.

We clearly have a data problem. I suspect everyone knows this. It's a data problem that is generated in part because the same act may be defined as a different type of crime in different jurisdictions. So those are simple definitional problems. We also have a data problem simply because the elements of the criminal justice system don't work for the Federal Government. They are in no sense obliged to supply us with data. If we want to know what is the variance in sentencing around the country for Robbery I, we may find court systems willing to provide us with that data, and we may find a lot of court systems that tell us it's none of our business.

In a national sense, the data problem in criminal justice means that we don't really know, can't really define, the basic systemic problems in a very quantified way. We have a feeling for where the system problems are, but we can't define them in a way that permits a quantitative evaluation to say, "Well, we have improved that problem."

Finally of course, one gets to the very basic question, the social question which asks how the criminal justice system, the police, the courts, corrections and citizen efforts, how do any and all of these operations affect crime rates in a jurisdiction? We know almost nothing about this. Yet these are really the basic mechanisms, the basic forces that a society can bring to bear in order to control crime.

We have in the first place a problem which is very poorly conceptualized, poorly defined in operational terms. Going beyond that, one again gets into the major data problems. But I think one can at least categorize the concepts. Criminal justice, through all its manifold efforts, is expected to bring about an effect of what is commonly called general deterrence in society. The fact is that because of the operation of the criminal justice system, a certain risk is involved in commiting an offense. That is, you are going to have to pay for it if you commit it and get caught. The idea of general deterrence is presumably that the operation of the criminal justice system keeps people from going out and robbing liquor stores. They don't do it because it's too risky.

We have no idea of the degree to which that concept is valid; and if it is valid, how do you go about measuring it? How can you decide in an evaluative sense whether more Draconian forms of punishment would in fact reduce the crime rate?

We know very little about the crime control aspects of the incapacitative effect. That seems very simple: when you put someone behind bars for three years for Robbery I, he may be doing nasty things behind the prison bars, but he is not out victimizing the public. However, we know very little about how much crime could be affected by a change in policy with regard to incapacitation--putting more people behind bars, putting fewer people behind bars, keeping them in the community. In point of fact, we don't even have very good statistics on how much time the average felon spends behind bars over the course of his criminal career.

There is another concept, and that is that once the offender has been involved with the criminal justice system, presumably it's had some kind of effect on his future willingness to commit crimes.

For a long time, we lived with an ideology of rehabilitation. It was a function of the criminal justice system to make useful citizens out of ex-offenders. That has very much come into question within the last year or so, partly on quasi-scientific grounds (there is very little evidence that this thing works in any wholesale sense) and partly, I think, because there is a tendency to move toward a more conservative philosophy with regard to the treatment of offenders.

These are the contexts in which we carry out the types of evaluations that we do carry out. Very briefly, our program is a grant/contract program. We, the Office of Evaluation and Office of Research Programs, which itself has a major evaluation program, are part of the National Institute within LEAA. The National Institute is set up and named in the law as the R&D part of LEAA. That means we (OE itself) are pretty far removed from decision-makers at the top level, that is, the administrators who make programmatic decisions, at least within whatever sphere of programmatic decision-making they have available to them under the Act.

What we do is essentially support major studies--usually of programs that are funded out of Washington. There is some money that is available to the administration for what are called discretionary grant programs--action programs designed in Washington, and open to competition. At the Administrator's request, we undertake studies of selected DF programs. These studies typically will take two or three years to do and cost half a million dollars.

We are also concerned with the much more basic problems, the problems whose solution could in the long run make a criminal justice evaluation a much more cost-effective undertaking. That is, we are interested and do support to a very limited extent a research program

that is taking a look at some of the basic problems like how you go about measuring a deterrence effect. How can you draw statistically valid inferences from police-recorded crime data?

Basically, that's where we are. I don't have a great number of success stories to tell you about the things we have accomplished so far. Maybe 10 years from now, we can have this conference again and we'll have some better examples.

MR. CAREY:

We have heard a lot there about how tough it is to get a handle on a problem that everybody understands. Now we are going to hear from the Environmental Protection Administration. Paul Brands is going to speak to us. He is speaking in the absence of Al Alm who is the Assistant Administrator for Planning and Evaluation. So it's good to have Paul here today, the Deputy Assistant Administrator, EPA.

THE AGENCY PERSPECTIVES PANEL (CONTINUED)

V. ENVIRONMENTAL PROTECTION

PAUL BRANDS, Deputy Assistant Administrator
for Planning and Evaluation,
Environmental Protection Agency

Thank you, Bill. I guess we are supposed to break around noon, so I will try to keep my remarks fairly short and hopefully relevant. Let me say that I am pleased to be here to share in some of the discussions of what Federal agencies are doing with respect to evaluation. Although I will not be able to attend all of these sessions, several people from our program evaluation staff are here, as well as others within the agency who are involved in evaluation; hopefully we will all come away somewhat smarter.

One word of background here. First of all, my office, Planning and Evaluation, generally is involved in the evaluation role in its entirety, if you assume a fairly loose definition of evaluation. However, within my office, we do have one division, the Program Evaluation Division, whose role in life really is to carry out evaluations, in the more traditional definition most of us give to that term.

To try to keep my remarks fairly brief, I'd like them to be guided by two criteria. First, rather than tell everything I know about evaluation in EPA, I'd want to emphasize those aspects which I believe are somewhat unique to us. Perhaps they are only unique as I see them, either because I don't know all that is going on in other agencies, or because I am somewhat biased in the way I view our impacts and our approach.

Second, because it's critical to those of us involved in the evaluation function, I want to focus on the process of doing evaluation, the organizational aspects, and how one feeds evaluation back into Agency planning.

The first thing I'd like to say--something that fits neither of the two criteria--is that there are two realizations which would logically argue that we ought to be emphasizing more the role and activity of evaluation within our agencies. All of us are confronted with a shortage of resources, and logic says, therefore, that you need to spend time and funds to try to find how to best allocate those resources you do have in trying to accomplish your task. And even in those instances when you have adequate resources, one can't just throw them at a problem and expect a reasonable solution. Again, one needs analyses, an evaluation in order to focus the efforts.

As I said, these two factors, I believe, tend to argue strongly that there will be more evaluation activities within the Government. I think, however, that at least one of those factors also argues that more evaluation efforts may not occur. As our resources get tighter and tighter, some managers within an agency begin looking fairly closely and longingly at those analysts who don't seem to be doing anything "constructive" (that is, the evaluators) and will try to get them involved in day-to-day operational activities. Certainly that is one concern I have within EPA.

Let me turn now to evaluation at EPA. We started the Program Evaluation Division in late 1973, staffed it up shortly thereafter, and I think we now have a pretty good program. Our intent was to develop an organization to try to determine to what extent the agency's

activities and programs as a whole constitute an effective, comprehensive attack on the nation's environmental problems. That is a very ambitious objective.

We look at the evaluation group also with the hope of their being able to provide fairly detailed information to our line managers, enabling them to better carry out individual programs.

In addition, we have drawn very heavily upon our evaluators to help us define operationally the agency's goals and objectives in our various programs, and to help the line managers look at those in quantitative, measurable terms so we can better assess where we are having some impact.

One area where EPA is perhaps unique is in the operational concept that we pursue within our Program Evaluation Division. We emphasize the relevancy of the evaluation the group is undertaking, the usefulness of the evaluation, and its potential impact on a program. We are not really interested in the ultimate report that may be written from the evaluation effort.

The second operational concept we have established is to work closely with the program office people as we carry out our evaluations. In fact, we have found (with the exception of only one evaluation) that by the time we have finished the report a large proportion of the recommendations in the report have already been implemented by the program office. We are pleased with this situation. I contrast this to what I have seen in several instances where the attitude of the evaluation people is to work in a secretive manner so as to come up with a startling report at the end--the idea being to have a big impact, not on what the agency is doing, but on the boss, by

showing him what great things have been discovered. In my view, that fundamental attitude or approach just does not result in an effective evaluation effort over time.

A couple of comments with respect to the audience targeted by these evaluations. First of all and certainly foremost, we do them for the internal managers within EPA, the actual program managers. The kinds of things that come up in these evaluations are recommendations with respect to resources, or organizational aspects; perhaps an evaluation will recommend a different mix of the subprograms which are being pursued in order to accomplish a particular programmatic objective. Or the evaluators may try to help define more precisely (or in more measurable terms) for the program people what their goals are or might be.

In addition, we are involved in carrying out evaluations which have been requested either by the Congress, by OMB, or by interagency groups addressing programs closely related to those of EPA.

Another point I want to touch upon is the organizational aspect of evaluation within EPA. The Program Evaluation Division is within my shop and under the Assistant Administrator for Planning and Management. This Division constitutes the focal point within the agency for major, comprehensive kinds of evaluations. Clearly, there are other groups within the agency who also carry out evaluations. We have a Management and Organizational Division within the Office of Planning and Management which undertakes evaluations, although these efforts are focused primarily on efficiency and organizational questions. We have the Program Analysis Division within our budget shop which addresses resource questions and evaluates primarily in the context of the budget.

In addition, our regional offices do a very limited amount of evaluation. At EPA, we have made a decision to develop an evaluation capability within the regions. I recognize the concern that was expressed in previous comments[7] and which must be kept in mind in pursuing this kind of course--i.e., that the regional evaluators spend their time doing regional policy analysis or evaluations which would be better undertaken at the national level. But in our case, we feel very strongly that the Regional Administrators are charged with carrying out a whole host of environmental programs. In the ten regions, we have a differing environment which we are trying to impact. Some of EPA's programs are much more relevant in some regions than others, and our view is that it's critical for that Regional Administrator to have some capability--some central capability within his region--that can, on a systematic basis, provide input to him as to which of the many national programs seem to have the most impact on the more severe environmental problems in his region.

With respect to this point, EPA has made a clear decision and we are pushing in that direction. We are still not where we would like to be with the development of this capability in the regions. We are finding that some Regional Administrators agree with our decision and are reallocating resources to carry out the evaluation function. But we still have a few who feel they don't need it.

A few more comments with respect to the staffing within the Program Evaluation Division. The formal evaluation group is not very large--in fact, it's only about 12 to 15 analysts. We have a number of approaches for augmenting that staff since no matter how

[7] See pages 40 through 43 above.

you cut it, that is a small group given the size of the agency, the magnitude of the dollar resources we are handling, and the severity of the environmental problem we are trying to improve.

First of all, we try to have the Program Evaluation Division take the lead in all our major evaluation activities. We augment that staff with some program people or with other analysts in the agency who know something about the particular problem or who have some sort of functional relevance to it (e.g., the organizational or budget aspect). We might wind up with a team of five analysts to address a particular problem, with from one to three of those coming from the Evaluation Division.

There are some real pluses to this approach, although I have debated this question with many people, in particular the GAO folks. From my point of view, I feel there are certain efficiencies associated with this approach, in that we can get "up to speed" much more quickly with a particular effort if we have substantial input and participation from the program people.

Secondly, because our ultimate goal is not just to write a report, but rather to implement our findings, we have found that program participation really facilitates actual implementation.

Finally, there is the important side aspect of enhancing the working relationships between the evaluation group with the program office as the evaluation effort proceeds.

One other comment with respect to staffing. We have followed the course of generally trying to maximize the use of in-house staff resources rather than going to consulting firms or others as some agencies do. One pays a price for not relying as heavily upon

outside capabilities in that you may not be able to take on as many evaluations as is desirable. In addition, you may not get quite as much expertise on the team as you may desire, at least in the beginning. But I think in our view, it is working out well because of the vast amount of programatic knowledge we develop within the Evaluation Division and the critical contribution of that knowledge to some of the other functions which the Division carries out.

That really brings me to the next point which, I think, from the evaluator's point of view, may be the most fundamental question of all. That is, after this evaluation is done, how does it get fed into the operational loop to make something happen because of it? It's the whole feedback issue. How do I insure an evaluation is fed into the Agency program planning cycle so that something happens because of this analysis?

Here again, I think EPA and the approach we have taken is somewhat unique. We have directly tied the Program Evaluation Division to the agency planning cycle; and we have done it in four ways. The Evaluation Division actually manages the four systems which are largely the guts of the process. The first system is program development. The Evaluation Division is involved in the actual writing and development of strategies for new programs. As you may recall, in the last two years the agency has had three new major pieces of legislation: the Drinking Water Act, the Toxic Substances Act and the Resource Recovery Act. In two of those three cases, analysts in the Program Evaluation Division were the key individuals in writing those strategies.

The second major system which the Evaluation Division manages is the MBO[8] system, which, incidentally has its pluses and minuses. In any case, to the extent that it has some meaning and impact within the Agency, the Evaluation Division manages that process and thus has an opportunity to insure that evaluations are considered.

The third area is the preparation of the annual agency guidance plan which lays out agency and program priorities, goals and the terms of measurement which both the headquarters and the regions are to gear their activities to in the coming year.

Finally, EPA annually ranks the many different objectives and programs which we have established to try to improve the environment in order to provide additional guidance in allocating agency resources. Again, management of that effort is carried out by the Evaluation Division.

We have thus tried to structure our system so that the people who are doing the evaluations are intimately involved with the major management systems within the agency, though insuring that we get maximum impact from the evaluation effort.

I have some notes here on some various evaluations that we have done. But, in the interest of time, I think I will skip them. Let me wind up by saying, perhaps in contrast to some of the earlier comments, I am fairly "upbeat" on evaluation, at least within EPA. Hopefully, I am still somewhat objective about where we are with it. I think generally it's seen by EPA management as a valuable, effective management tool. I think we are committed to its continued use and growth.

---

[8] Management by objectives.

I should point out, however, that we are not without many of the problems everyone in this business is confronted with--e.g., trying to measure effectiveness, and attribution when various levels of Government are involved. I think the other aspect that troubles me sometimes is that we do not have enough time or resources to ask some of the very fundamental questions which a true evaluation should; for example (with respect to our agency) what programs are really cleaning up the environment from the health and the ecological viewpoints? Perhaps over time we will get closer to addressing these types of questions.

Thank you.

VI. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

MR. CAREY:

We have heard from Dick Linster of LEAA and Paul Brands from EPA. I wonder if the audience would have a couple of questions that they'd like to get in here. I'd be very much surprised, for example, if Jerry Caplan didn't want to ask a question.

PARTICIPANT:

Jane Woodward, the Urban Institute. I have three questions or three interrelated questions. One I'd like to direct toward John Evans, one toward Dick Linster and one for whomever would like to take it.

The real question about Federal-level objectivity that I'd like to ask is this: is it not really the case that Federal policymakers and program managers have as much invested in their programs and policies as state-level policymakers and program managers, such that does the Federal contracting process really guarantee greater objectivity than the state-level contracting process? They both are contracting processes. That question was for John Evans. This one is for Dick Linster. LEAA has conducted a great many evaluations at the Federal and state and local levels. Therefore, could you address whether you have found there are greater levels of objectivity in evaluations at the several levels?

The third question is: haven't we really been talking about one type of evaluation, really the kind of evaluation that leads to generalizable knowledge? And is that the only kind of evaluation we are in the business of conducting?

MR. EVANS:

The point I was trying to make about objectivity before was that the competence, utility, and objectivity of an evaluation is not necessarily related, as Bill quickly pointed out, to its geographical location. But it is related to the involvement in program activities and the responsibilities for program activities which people have who carry it out. What I am saying is that within some Federal agencies, one can find that the evaluation function is highly decentralized so each program or each bureau is responsible for conducting evaluations of its own activities. The point I was trying to make is that that is an inherently unwise situation for setting up competence and productive evaluations, where the end results are intended to assist the overall decision-maker or agency head to make comparative and objective program judgments. You put the head of a program in what I think is an impossible situation when you ask him to evaluate his own efforts. What you frequently find is that if he is really a devoted and competent program director, he "knows" his program is good, and he is not about to spend $500,000 on a study which asks if it is any good. He would much rather spend that on management improvement or on the program itself. That kind of a situation is to be contrasted to one where you purposely set up a separate organization, staffed by technically competent evaluation people who have no program responsibilities, and therefore no extremely parochial commitment to those program activities. So you try to combine evaluation technical competence with non-commitment to the program.

It's interesting that that argument can be carried several steps further. One can say, "Well, if that's true, shouldn't the evaluation of Federal programs be outside the Federal agencies altogether--in the GAO perhaps, or other outside institutes?"

That is logically where you are driven. At some point you simply have to make a decision between maximizing objectivity and technical competence on the one hand, but not getting into a situation where the evaluation activity is so remote and so removed from the policy and budgetary mechanisms that the character of the evaluation and its results are likely to drift into less relevance. I don't want to say any more on that because we have other matters to pursue.

MR. CAREY:

The Chair rules that the question has been answered. Next, Dick Linster.

MR. LINSTER:

I am not sure this is a direct answer to your question, but I think objectivity is only one of the criteria by which one presumably would judge whether the resources you put into an evaluation were well spent. Within the LEAA program, there are a lot of evaluations that have nothing to do with Washington. They are done at the local level sponsored by state planning agencies. I don't think that of the good ones, the ones that are technically sound, I don't think there is a real question of objectivity. I think the objectivity is just as defensible there as it is for anything we might sponsor. Maybe that is enough.

MR. STROMSDORFER:[9]

You seek information at many different levels when you are attempting to devise and operate programs. You seek generalized knowledge and information about the state of the nature out there with respect to health and occupational safety, for example. You seek more narrowly focused information about the political impacts

[9] Agency Panel Member Ernst W. Stromsdorfer, Deputy Assistant Secretary for Research and Evaluation, Department of Labor.

of the behavior you are attempting to encourage. You seek very particularistic information about the internal management and efficiency of a program. You seek all these kinds of information. They all go into generating a body of knowledge and understanding about how a program operates.

MR. CAREY:

Is there a question from the Republican side of the room?

PARTICIPANT:

Walter Bergman with the IRS. The question is to Dick Linster. Could you just address yourself a little bit to the area of white collar crime and to what is being done with regard to research in that area and not just the local and state, but also the Federal effort in criminal justice?

MR. CAREY:

In a couple of short sentences.

MR. LINSTER:

Jerry, would you like to take the microphone? I know we have a program in the area, but I know none of the details.

MR. CAPLAN:

Very briefly, we are doing some interesting things. One program will be studying ways to minimize frauds against governmental benefit programs. We also have a long-range grant with Yale University for a five year project, two years of which have been funded. The project is a multi-disciplinary study of white collar crime. Because this is almost virgin soil in terms of research, the first year has been a planning effort. They are developing an extensive research agenda including a study of Federal regulatory agencies and the

process of referral of agency cases for criminal prosecution. To the extent that these have been touched upon, the questions raised have related more to such issues as whether there should be an adjudicatory hearing or an administrative hearing. Another program is looking at corruption in state and local licensing and regulatory agencies. This will include government contracting (the kind of thing that involved Governor Agnew), licensing at the municipal level where we suspect there are very interesting patterns of corruption and non-compliance, housing, all the areas where little research has been done. Our research will attempt to delve into the nature and patterns of corruption in these areas. At the same time, we have some more conventional efforts under way on shoplifting and employee theft, which I view as more manageable research but less exciting.

MR. CAREY:

Neatly done. Now I am going to have to stop the questions. We have had at least a little ventilation. Before we break for lunch, I want to get one more speaker through. That will leave us still with four after lunch.

I am going to turn to the Bureau of Indian Affairs, Bob Hemmes; but I am reminded of one noc.day in LBJ's administration when I was in the East Room of the White House; they were having a big celebration for bill signing. A bill had been passed in the area of Indian affairs, and all the Indian Chiefs who could be identified were there and a great many other people. The room was packed. When the President came in, he decided that he was going to take full advantage of it; and he threw away the speech that the speech writers had all cooked up for him, and he got in there with feeling and emotion. As he built up his momentum, he'd look out and say, "Now, Willard Wirtz, I am telling you right here in the presence of all these people that I want you to do the following four things for

these wonderful Indians. John Gardner, you haven't been doing enough." And he went after John Gardner, and Sarge Shriver with the same thing. This went on and on. He was really performing.

I was standing in back in the corner with Joe Califano and a couple of others, and Joe whispered, "Get somebody to pull the plug on his mike. He's giving the country back to the Indians."

Well, I don't know whether the Indians have got it yet or not, but let's hear from Bob Hemmes, Chief of Planning in the Bureau of Indian Affairs.

VII. INDIAN AFFAIRS

ROBERT A. HEMMES, Chief of Planning,
Division of Transportation, Bureau of
Indian Affairs

I am going to take advantage of Jerry Caplan's suggestion to gain sympathy for my story. In fact, I'll go him one better. Instead of a self-deprecating remark, I am going to introduce a non-self-but-deprecating remark which will also make another point, that all of us think our job is the toughest.

The remark was published in the newspaper. This is Monday's (November 15, 1976, p. 8-2) ever-popular Washington Star. I am going to quote the first paragraph verbatim.

"The Washington headquarters office of the Interior Department's Bureau of Indian Affairs is wildly mismanaged."

Evaluation is certainly the new game in town, and it reminds those of us who have been here more than just the last year that this is only one of a sequence of games called value engineering, zero defects, benefit/cost, cost-effectiveness, systems analysis, PPBS, MBO--and now evaluation. This is a time of change, and people are jockeying for positions in the new administration. They don't want to be left out.

But these "games" have a great commonality. They all have a perspective on the same problem: What are we trying to do, and how well are we doing it?

The first requisite of evaluation is answering the question: what is it we are trying to do? This is usually a hard question. The laws, as was pointed out, are vague and general. That is, the price paid for the consensus needed to pass a law is to state something that is vague, general, virtuous, desirable, pointing to a direction in which to go with which everybody agrees. It might be something like "Improve transportation." There is actually an Indian law which says, "Support civilization." Or we might have something like, "Eliminate hunger" which I am sure is the primary thing on all your minds at this noontime. Or other vague and general things.

In order to get a handle on these vague and general things, which are usually called "goals," it is necessary to break them up into smaller pieces, i.e., subgoals, sub-subgoals, sub-sub-subgoals, etc. In order to avoid getting confused with all the "sub-subs," I prefer to break up goals into an (arbitrary) hierarchy: goal, objective, mission, purpose, task, job.

Let me tell you a little about the Bureau of Indian Affairs and where we are on evaluation. As someone so delicately put it when I walked into the room this morning, "Your interview[10] was refreshing because you were the only one who said you didn't know what you were doing." Now I had some help in arriving at this position.

---

[10] See Appendix II to the present Volume.

Historically, the Bureau of Indian Affairs and its progress have been conditioned by a series of events, policies and laws. The Bureau was created in 1824. In 1830 Congress passed the Indian Removal Act which began the relocation of tribes from the East Coast to remote Western reservations. During 1870-1886 Federal Indian policy, administered by the Army, completed the relocation of Indians to reservations and began giving them food and clothing rations. Stories persist that the Army tried to exterminate Indian people by typhus-infected blankets, starvation, allowing disease to go unchecked, and shooting. In 1887, the Dawes Severalty Act broke up the reservations by providing individual land allotments, and opening the balance of the treaty reservations to non-Indian settlement. All Indians received citizenship and the right to vote in 1924 (although some Indians were already citizens by virtue of their treaties).

The Indian Reorganization Act of 1934 ended land allotments and provided for tribal self-government. This is the act that took the "chiefs" out of the tribes and put them in the Civil Service as Division and Branch "chiefs." Indian leaders became known as "governors" or "chairmen." In 1953, House Concurrent Resolution 108 called for termination of Federal trusteeship over Indian tribes and their affairs and property. The Menominees of Wisconsin was the first large tribe to be terminated. Also in 1953, prohibition for Indians was repealed. In 1957 the Bureau began relocating Indians off the reservations to make them "part of the mainstream of American life." The termination policy was reversed by President Nixon's special message to Congress in 1970, setting forth the Indian's right to self-determination without threat of termination. The relocation policy was reversed in 1972. The Menominee termination was repealed in 1973.

Self-determination for Indian tribes is now assured by the Indian Self-Determination Act (passed January 5, 1975), which provides the opportunity for Indian tribes, upon request, to take over any program administered by the Bureau of Indian Affairs along with the funds to run it--probably the most enlightened of all these historical policies.

This historical back and forth, and crisscross of policies has left us not only with checkerboard Indian lands, but checkered policies. We don't have a long history of building up a rational and systematic evaluation capability, nor do we have much of a historical data base for long-range analyses.

But we do have opportunity, and I think the opportunity we have may be the greatest in the Federal Government because the Bureau is unique in two respects. The first is that our constituency, a very small minority group of 500,000, has a unique relationship with the Federal Government. It has a claim to sovereignty based on treaties with the United States of America. That is something to consider. No other minority group has that sort of status.

The second unique aspect of the Bureau is that we have all of the functions of the Federal Government. We are a Federal microcosm. We have offices corresponding to Federal departments for trust responsibilities and services, business development, financial assistance (grants, loans, and loan guarantees), job placement and training, transportation, law enforcement, tribal government services, social services, housing, planning, schools and education, and numerous support services. That is also something to consider. The Bureau of Indian Affairs has the opportunity to develop, not only evaluation, but all the analytical techniques and methodologies which could be exemplary and serve as a model for all of the Federal departments.

Evaluation, without goals and objectives, is not helpful. I am in the Transportation Division of the Bureau, and we are still at the roadbuilding stage. We don't have any automated vehicles or rail lines or even buses, but we are building roads. What good are roads if they don't go anywhere? It's just like evaluation. What good is evaluation if it doesn't go anywhere? Evaluation doesn't tell the whole story. Evaluation is for a purpose, so it's a part of something that is at a higher level of abstraction.

On the national level, we have the laws--vague and general laws which express national goals and which are determined by the wisdom, judgment and experience of the nation's leaders (see Figure 1). They name things. They are on a nominal scale because they name things just as all of us in the room have a name. Names are useful to talk about something. They are useful for information retrieval, but they really don't have any meaning. Your name might mean something

PROGRAM EXPLICATION

ORGANIZATION	CHARTER	WHAT	ANALYTICAL TECHNIQUES	SCALE
NATION	LAW	GOAL	WISDOM, JUDGEMENT AND EXPERIENCE	NOMINAL
DEPARTMENT	CODE OF FEDERAL REGULATIONS	OBJECTIVE	COST/EFFECTIVENESS	
BUREAU	MANUAL	MISSION	BENEFIT/COST	ORDINAL
DIVISION	ORDER	FUNCTION	BENEFIT MINUS COST	
BRANCH	MEMORANDUM	PURPOSE	Δ BENEFIT/Δ COST	INTERVAL
SECTION	VERBAL	TASK	ECONOMETRIC MODEL	
PERSON	POSITION DESCRIPTION	ASSIGNMENT	OPERATIONS RESEARCH ALGORITHMS	RATIO

FIGURE 1

Robert A. Hemmes, Agency Perspectives Panel, Symposium on the Use of Evaluation by Federal Agencies, MITRE/METREK, November 1976.

from its origin or derivation, but it's just a designator or symbol. The laws are much like that. They name some desirable goal that we are all pursuing. But when we come to the next level, the department level, we have a subgoal or "objective." The objective says how far to go down this virtuous path. The evaluation technique that is associated with the objective, I suppose in its broadest sense, is called cost/effectiveness. If we can't measure effectiveness, we name it effectiveness. All of us can make up something to represent effectiveness or how we feel about it.

Generally we can measure cost. It isn't always easy because we have the problem of cost allocations. Now we are beginning a process which I am trying to promote within the Bureau of defining what we are trying to do and where we are going. This process I call "explication." I have begun the framework of the explication by going from the national level to the department level, by going from the goal to the objective and on down, as you can see in Figure 1.

The columns are, first, the organizational entities, second, the written charters that enable them to operate. The third column is a "what" column for lack of a better word, that is, what they are doing. Then, there is "evaluation," and finally, scaling. Evaluation has an implication of measurement.

There are four ways to measure something. Scaling techniques in my context are from an article by S. S. Stevens written for Science Magazine in the '40's. It has been widely used in the behavioral sciences because they have a tough problem too.

A lot of engineers who grew up with the dimensions of "mass," "length" and "time" throw rocks at people in the so-called inexact or soft sciences because they don't know how to measure anything. The problem is, they have harder things to measure. At the second level, the Department level, the laws are written and published in the Federal Register and codified in the Code of Federal Regulations, the CFR.

I guess it's pretty widely known that the Exeuctive does rewrite all the legislation. It's becoming even more widely known that, in this rewrite, omissions from the original legislation occur, and additions to the original legislation also find their way into the text. The reason for CFR is obvious. The laws, being vague and general, don't say what to do. They don't, in fact, say much of anything. So the explicative process is being carried out, first of all, by the CFR. The CFR says how the law is going to be administered, but not too specifically.

Then we get down to the Bureau level, and the Bureau has a manual. You all have counterparts of a manual. Looking at the evaluation column, you see that cost/effectiveness measurement now becomes possible; as we get down into the smaller units and we improve the scaling techniques, we can measure effects in dollar benefit terms. If so, we can form the ubiquitous benefit-cost ratio invented and pioneered by the Corps of Engineers as a result of the Flood Control Act of 1936. The Corps deserves a lot of credit for also pioneering the five ways to cheat in the benefit-cost ratio: lie about the cost, lie about the benefits, use an abnormally low rate of interest to discount the benefits to present worth, and extend the economic life. The fifth one is highly imaginative--find the worst way to do it, and count the cost difference between the way you want

to do it, and the worst way to do it as a cost saving. The benefit-cost ratio is a two edged sword, however. It serves the Corps well because it puts them into a nice coincidence with the wishes of the Congress.

Moving along, (fourth level, Figure 1) many people have suggested that benefit minus cost might be a better measure because benefit-cost is a go, no-go, test. It's a gate. You cannot rank by a benefit-cost ratio because ranking by ratios is meaningless. All you can do is divide the project into two lists--those with a favorable benefit-cost ratio greater than one; and those with an unfavorable benefit-cost ratio less than one.

If we want to go into ranking, we are moving from the nominal to the ordinal scale in the last column. To get to the ordinal scale so you can order something, you have to use the incremental benefit-cost ratio. Delta B over delta C can order. We are moving to more powerful scales now down this hierarchy. Down below the Bureau level, we have the Division, and Branch; these are arbitrary; but most organizations have a division, a branch and a section.

We have the explication down to the position description--what it is we are going to do. As we get down to the jobs that can be handled by one man, perhaps they are amenable to more powerful evaluation techniques provided we can move down to the corresponding scales. If we move from nominal to ordinal, we can say which precedes what.

I used to work for the Department of Transportation, and the Office of the Secretary used to ask me to submit my list of R&D projects by priority. What they were saying was, what's first? What could I do but put them in alphabetical order? They didn't give me any ordering criterion. In order to move to the ordinal

scale, you must face up to the ordering criterion. This is difficult in social programs because it can be shown (and in fact, Kenneth Arrow showed), that social choice is intransitive.[11] If A is preferred to B and B is preferred to C, you cannot conclude that A is preferred to C because you can construct a counter example where C is preferred to A. So it's circular.

A lot of Government programs get off to a good start and then run in circles when they get down one level of abstraction to the ordinal scale.

If we were to establish some kind of unit for utility or usefulness, even an arbitrarily scaled unit, then we could move to a more powerful scale called the "interval scale." Using the interval scale, we can make statements about the difference between A and B. A minus B--that is a more powerful statement.

A ratio scale introduces the notion of a zero or a data plane, and that enables you to make a statement like A is so many times bigger than B. That is really what we want to know about Government programs--where they stand in the hierarchy, whether they are amenable to evaluation, what is the most powerful type of evaluation we can use on them and how can we come up with a priority list that is meaningful?

There are lots of difficulties. One primary difficulty is that when you get down to ordering you have to have an ordering criterion. You have to order on a single principle. I can't say what is first by age and weight and alphabetically and salary. I have to give you

---

[11] Editor's note: See Bauer, Raymond A. and Gergen, Kenneth J., The Study of Policy Formation, The Free Press, New York, 1968, pp. 60-61.

one or the other, but yet we often want to know all the attributes of a person or project so we have a "vector of attributes." We might have four numbers, A, B, C, D, with parentheses surrounding them. That's a vector. That is the state of the system. How do you order a vector? Well, there are several ways to order a vector. One is lexicographic ordering which is kind of alphabetizing and not too useful in, to coin a word, "prioritizing" Government programs.

Another way to order vectors is a geometric distance--the square root of the sum of the squares. That doesn't always work, it collapses everything into a scalar and all the information in the vector is lost. A Miss America Beauty Contest is an example. You have three dimensions in Miss America--the bathing suit, the evening gown and the talent. Suppose you want to hire her for a night club act. Maybe Miss America was Miss Colorado because of the way she looks, but maybe the best singer was Miss Utah. So you don't know who Miss America really is until you know her future objective.

Think about the same idea in a Government program. What is its objective? Whose principal interest is involved? There is where the political process comes in in establishing the criterion for evaluation.

MR. CAREY:

I find myself almost speechless after that. We have a problem of choice. Do we eat, or do we talk? Is there anyone in the room with an irrepressible question for Bob Hemmes? Now is your chance.

We will declare the morning's session over and go to lunch. Thank you all very, very much. See you after lunch.

LUNCHEON ADDRESS

EVALUATION AND THE BUDGETARY DECISION-MAKING PROCESS

TONEY HEAD, Acting Deputy Associate Director
for Evaluation and Program Implementation,
Office of Management and Budget

MR. BENINGTON:

Our luncheon speaker today is Toney Head of the Office of Management and Budget. He and I met before lunch and realized our paths have crossed a number of times. In discussing the meeting at hand, Toney said that he thought there are some very tough questions that should be asked. For example, he said, "Do we need LEAA?" Well, don't ask too tough questions.

He also asked some questions about work that my company does for the Government as to whether we really are assisting in the best role possible; whether we are honest; whether we are tough; whether we in fact follow through with a lot of the rhetoric. He became very specific, not only with respect to our LEAA work (where we are doing a splendid job), but with respect to other organizations. So I figure that Toney and I are now very close friends. Let me introduce him.

Toney Head is now Deputy Associate Director for the Evaluation and Program Implementation Division of OMB. He is responsible for the development and implementation of Government-wide evaluation policies, for the administration of the Federal Advisory Committee activities, for the promulgation of management improvement policies and the assessment of agency efforts to improve management--just the things he and I agreed all of OMB should be doing. He has been

there since 1970. He is a graduate in management from Maryland and Syracuse. He has a very wide experience, including having worked in the Department of Defense and with the U.S. Army, where he learned what great management techniques are.

MR. HEAD:

Thank you. I am delighted to be here today to have this opportunity to comment on OMB's role in evaluation. Before I begin, I would like to speak very briefly to some of the major problems that we have in evaluation today. Although they are not in priority sequence, in each of these lies a major cause of evaluation failure.

Number one, there is an overall lack of clarity and consensus on the objectives of Governmental programs. Too often these objectives are Utopian and vague. In many instances, programs are stated in such convoluted prose that there is no way of determining whether their objectives are accomplished or not except for intuitive feeling.

The program legislative authorization process itself involves compromises among oppo ing positions. These, in turn, are reflected in ambiguous program objectives.

The second area relates to poor management of evaluation findings by agencies. Decision processes do not use evaluation results, regardless of how good they may be. In many instances, evaluation results are not utilized at all. Those responsible for evaluation are too often not at the top, so they are not the policymakers formulating the decisions. Hence, evaluations are often regarded as irrelevant.

Another problem is that evaluators are frequently given other tasks. Not enough resources remain for evaluation management. In many instances evaluation staff resources are diverted to crisis management and planning. Agency RFPs.[12] contain too little information on what the agency wants; too much is left to guesswork by the contractor.

The third area involves a lack of incentive for Governmental managers to critically evaluate their program activities. Unfortunately, in the Federal Government, we do not have the income statements that they have in private enterprise. This, as you know, forces management in private enterprise to eliminate those activities and programs which are not contriubting to whatever the program or organization is doing. We do not have those kinds of "forcing elements" in Government.

Bureaucrats or Governmental managers get attention if they build large organizations or if they start new programs, not if their programs are effectively run. Also, too many program managers allow their personal reputations to ride on program successes. Regardless of how ill-defined a program may be, personal reputations are attached to its success.

The fourth area is the complexity of most Governmental programs. This makes cost-effectiveness and program impact analysis difficult. Many of these programs affect all parts of an industry or social condition.

Efforts are often divided up among different approaches. In many instances, no single organization or unit within an agency has overall control of all the various approaches that may be used.

---

[12] Requests for proposal.

There is much ad hoc responsibility, which in many instances, is not documented. Organizational units handle several different programs within the same organization.

Another point relates to the theoretical methodological deficiencies in evaluation techniques. Many of us in the Federal Government say that we can't evaluate certain programs because the methodologies and techniques have not been developed. Although to a large extent, we use this as a crutch, there are major deficiencies in this area.

First, measures of effects in many instances are lacking. A statistical approach requires considerably more cases than can often be afforded. Data sources are undependable over time. In many instances, we have established programs, we have implemented them, we have administered them over a period of time without even considering what data is needed to determine the results or measure the results against the objectives of the program.

The rational decision process assumed by most evaluation efforts is not followed in practice. However, in many instances, there is no adequate alternative model.

Evaluations generally attempt to do too much. They try to get dramatic, overall, law-of-nature results when the program could not conceivably have such effects, instead of focusing on getting limited but practical information about one or two basic program assumptions.

In examining OMB's role in evaluation, one must look at the President's responsibilities and how OMB supports the President in meeting those responsibilities. Under the Constitution the

President is charged with insuring that the laws of the country are faithfully implemented. Embodied in this charge is the need to insure that resources are utilized in an effective and efficient manner and that Governmental resources are applied to accomplish the intended results of the laws that he is charged with faithfully implementing. Evaluation is part of that responsibility.

In meeting this responsibility the major arm of Government that is used to support the President is OMB. The agency was established in 1921 to help the Chief Executive prepare the national budget. As most of you know, prior to that, each agency submitted its budget separately to Congress. There was not a national budget. Since 1921, several laws have been enacted that have augmented the management responsibility of the Director of OMB as well as of the President.

Then in 1970, Reorganization Plan No. 2 was announced. Management responsibilities which to that date, had been given by law directly to the Director of BOB[13] were now transferred back to the President. Then the President redelegated those management responsibilities to the Director of OMB.

To summarize these twofold responsibilities in brief, they are the following: first, to develop Government-wide management policy and second, to monitor and evaluate the efforts of agencies in meeting their management responsibilities and to report the results to the President.

Within that responsibility, of course, is OMB's evaluation role. Before I get into the discussion of that role, you should consider certain basic assumptions.

---

[13] Bureau of the Budget, pre-1970 name of the present Office of Management and Budget.

First, the management of programs is an agency responsibility. The management responsibilities which we said a few moments ago were delegated by the President to the Director of OMB do not include the administration of Federal programs. The management of Federal programs is the responsibility of the departments and agencies.

A second assumption is that OMB will meet its responsibility by providing Government-wide policy guidelines and through selectively monitoring and evaluating the efforts of agencies. As I comment on OMB's role, you will find it will fit in those parameters. Basically, there are four aspects to OMB's evaluation role.

OMB's number one charge is to provide Government-wide policy guidance. We have done that in Circular No. A-11 and, to some extent, in Circular No. A-44.[14] As for our second charge, to monitor and selectively review agency evaluation systems, I will comment further on that as we go along. Thirdly, we incorporate program evaluation concerns into the budget process whenever possible. Finally, we are to provide leadership and direction to the Government-wide efforts to improve the conduct and practice of evaluation. Basically, those are the roles of OMB in the area of evaluation.

I would like to briefly describe the activities we have undertaken during the past two years. I might say that if you have to rate us on our past performance, it would probably be marginal or perhaps somewhat

---

[14] Circular No. A-11, Revised, dated July 16, 1976; Subject: Preparation and Submission of Budget Estimates.
Circular No. A-44, Revised, dated May 24, 1972; Subject: Management Review and Improvement Program.

better than that. However, we have done some meaningful things. One example is a survey of evaluation activities of major agencies in Government which we conducted in cooperation with the General Accounting Office. We identified organizational structures, each of these agency's concepts and approaches to evaluation, and the estimated costs of those evaluation activities.

We also established an inter-departmental panel of senior evaluation officials, usually at the Assistant Secretary level. The major purpose of this panel is to discuss issues and problems of common interest in evaluation.

Another important activity is our provision of technical assistance to agencies. This has probably been one of the most meaningful activities that we have participated in.

Additionally, background papers have been developed which discuss problems associated with planning and management of evaluation projects. These background papers have been circulated throughout most of the major agencies in Government. Agencies have commented on them in draft form, and they have since been published for a Federal audience.

Within OMB, we drafted an Evaluation Circular which, as of this date, has not been signed. This has been circulated and coordinated with all agencies. In many instances, agencies have made major contributions to that circular.

In-depth assessment of selected agencies' evaluation activities is another important activity which we regard as a very meaningful effort. Oddly enough, we found that some of the major departments

have no central evaluation capability. Although they have major responsibilities for the implementation of national programs, the evaluation capability, if any, is under the individual program manager. In many instances, this is designed merely to meet his day-to-day requirements for implementation and administration of the program.

In some cases, we discovered that agencies have been established for two or more years and still have no evaluation capability whatsoever, either by the program manager or by a separate unit reporting to the agency head. You might think that is unusual, but the fact of the matter is that this situation exists. In other instances, we found there is a separate evaluation activity in the agency, but it is at the lowest echelon of the organization. There is little or no possibility for any of these evaluations to impact on the decisions made within that agency.

Another area in which we have done some work is a comprehensive analysis of evaluation training needs of Federal executives. We have worked on this in conjunction with the Civil Service Commission; the results will be published shortly.

It has been important for us to maintain liaison with the Legislative Branch. We have worked with the GAO in encouraging certain subcommittees within Congress to do a better job in identifying what many of these programs should be doing. GAO, in the past five years, has made specific recommendations to Congress along this line. These recommendations stem from the fact that, in many instances, GAO has gone into an agency to evaluate a program and has identified the objectives of that particular program, only to find that the agency does not agree with those objectives. After examining the legislation

they have often found it to be ambiguous. GAO and the agency fail to reach agreement and must trace the legislative history in order to try to identify the intent of Congress in the establishment of that program. This difficulty has been pointed out to a number of subcommittees in Congress. Congress has been urged to do a better job in identifying the specific objectives of the programs.

I might comment briefly on some of the current efforts of OMB.

With respect to evaluation policy, we have a draft circular that has not been signed, but we have not given up. We are moving forward in a number of ways. There is already policy guidance in A-11 which we will augment by either issuing a separate circular or including evaluation policy in an overall management circular which replaces A-44. We are convinced policy guidance in this area is needed and we expect OMB to eventually promulgate this policy.

Another initiative within OMB involves technical assistance for the budget examiners. During the budget process the evaluation unit has placed very high priority on continuously working with each budget program examiner. There is also technical assistance to agencies. We are currently doing a great deal of work with the Bureau of Indian Affairs, the Veteran's Administration, and some of the other agencies.

We are now examining strategies by which we can make evaluation a special component of the budget process. This can be done in a number of ways. One would be to better utilize current strategy by working on a more continuing basis. An evaluation specialist could collaborate more closely with the budget examiner during the complete budget process. Another possible strategy would be to make a special component of the budget process the discussion of evaluative issues. There are several others that we are considering.

Another ongoing interagency effort is the development of an evaluation network system. We are trying to identify the specific evaluative information needed by OMB which would be most useful to the Executive decision-making process. This includes evaluation information needed by examiners in the budget formulation as well as information needed by the President or the Domestic Council.

Finally, we are studying the kinds of evaluative information that may cut across agency lines and which may be needed by several agencies. We will not know the results of this undertaking until we can complete the identification of the information requirements I just mentioned.

In summary, I would just like to state that OMB can provide the policy guidance which should reinforce the management framework within which agencies can develop more effective evaluation systems to better support their decision-making process. In the final analysis, however, it is up to the agencies themselves to conduct meaningful evaluation activities and to insure that evaluation results are considered in the decision-making process.

I am open to any questions which you may have.

MR. GRANDY:

I wonder if you could amplify a little bit on this evaluation network system that you mentioned as being a way to get the information flowing in. Could you describe that a little more for us?

MR. HEAD:

One of the things that we have identified is that our OMB examiners, in many instances, are making recommendations to the agencies. To some extent, these go beyond recommendations concerning the funding levels of certain programs. Too often, examiners do not really know the impact of those programs although they need this information. This need has to be identified prior to the budget hearing. In many instances, it must be identified one or two years in advance. We know that, in general, this is not being done. Certain kinds of evaluative information are needed within OMB, but at this time, we have not identified the specifics of these information needs. Some of these evaluation information needs cut across agency lines. Many of the evaluations conducted by HUD, for instance, are directly of interest to HEW and vice versa. This is recognized, but we do not always know the kinds of information needed.

This has been an interagency effort and not just OMB looking at it alone. We have a person from the Department of Health, Education and Welfare (HEW). We have another person from the Department of Commerce. We have individuals from two or three other agencies who are participating. We think this has been a very meaningful effort.

Let me make one other comment. We also know that the General Accounting Office is gathering all kinds of information on evaluation. We are working with the General Accounting Office to make sure we do not end up requesting information from agencies that they are already gathering. Hopefully, they are doing the same.

PARTICIPANT:

Would you identify what this technical assistance part of the program is, the kind of activities included?

MR. HEAD:

Let me illustrate by telling you how we have worked with the Veterans Administration. I believe there was a law passed in 1975 which requires the Veterans Administration to conduct an impact evaluation of all programs on an annual basis. This is nearly an impossible task. At that time, the Veterans Administration had no separate evaluative unit which could conduct impact evaluations. We worked with the Veterans Administration in setting up an evaluative unit and developing some kind of a strategy under which they would plan to conduct evaluations of certain programs on an annual basis. Another example is our work with the Commodity Futures Trading Commission which was established about a year and a half ago. About a month ago it was discovered that they did not have an evaluation unit; previously there was little, or perhaps no interest in evaluation. The commission recognized that an evaluation capability was needed and they came to us; we are working with them on establishing an evaluation system.

PARTICIPANT:

Is that properly called technical assistance or is it really an assertion of higher management's preferences, would you say?

MR. HEAD:

We did not go to these agencies and ask that they establish evaluation capabilities or that they change their system. They came to us and said, "We have a problem." In the case of the Commodity Futures Trading Commission, they said, "We have a problem in evaluation. We haven't addressed it, and we want to know how to go about it."

We gave them certain suggestions, but we also referred them to other small Governmental organizations that had similar problems. They did not work with us alone. Although they have not yet arrived at what it is they are going to do, when they do, it will be their decision, and not ours.

PARTICIPANT:

Jim Robinson, Labor. Isn't one of the main problems with OMB's overseeing an evaluation program that most of the weapons that OMB has to work with are basically negative rather than positive? What I am thinking of is, faced with the fiscal constraints we have been having over the past five years and are likely to continue to have over the coming years, evaluation becomes much more an exercise in, "Which program can we do away with to free up new money so we can start another initiative?" or else "How can we straight-line a program to free up more money?" If you are really looking at evaluation from that point of view, regardless of what happens in A-44 or another OMB circular, if all the promotions are given to a guy who tears a program down or puts one out of business, rather than one who builds one from the bottom up, how do you really have the capacity to institutionalize evaluations?

The other part of that is what sort of accountability is OMB willing to stress in its evaluation program? Are you willing to identify managers who have not evaluated successfully and whom you have removed? Are you willing to identify managers who have evaluated successfully and see to it they have been moved up the hierarchy to teach a lesson to other people that it pays off to evaluate? If you really want evaluation to work, you have to make sure you approach it from the positive point of view of rewards, and some of your evaluation is going to cost money and some is going to save money.

MR. HEAD:

Let me begin by answering your first question, Jim. One of the initiatives I mentioned refers to our objective of having an evaluation specialist working together with the budget examiner. When you were in OMB, unfortunately that was not the case. You did not have an evaluation specialist advising you regarding the kinds of evaluative questions you should be raising with respect to programs for which you had responsibility. We are trying to get away from the particular environment that you just described. Some of these initiatives that I have mentioned are efforts in this direction. I think that the budget examiner is just as interested in good management as the person at the agency level. I will admit that, in many instances, his focus is very narrow and he is directly concerned with funding levels of a particular program.

During my discussion of problem areas I mentioned that there is a lack of attention both to the results of programs and to using this kind of information in determining their worth. This applies to establishing funding levels for the program, discriminating between programs which might be eliminated, and programs to be maintained at either an increased or a lower funding level. Evaluation results are needed to assist in making these decisions.

While I cannot now state that we are using this information in an effective manner, I can merely say that we have initiatives ongoing that will improve our use of evaluation information and will identify specific needs for evaluation information a year or two years in advance. In the future this would allow the agencies to gather this information for budgetary decisions.

CONTINUED

1 OF 4

Now, you ask what is OMB doing to promote those managers or demote those that are not doing a good job. Please keep in mind that the chief responsibility for management lies with the agency. Any recommendations on the part of OMB to demote or promote managers would circumvent an agency's decision-making process. There are other things OMB can do that will give extra recognition to those managers. One is a Presidential Management Improvement Award. The Civil Service Commission has been encouraging agencies to recognize those Federal managers who excel in administration of their programs. Another method would be to single out certain managers for special recognition of individuals responsible for more limited initiatives as well as for major efforts.

I apologize that this question must unfortunately be the last question due to time constraints. It has been a pleasure speaking to you.

MR. BENINGTON:

Thank you very much. I promised Toney that I'd protect him, not from you, but from his calendar. He has to get back to his office, and now we'll go back to MITRE. Thank you.

MR. GRANDY:

In the afternoon part of our program, although we are falling behind our expected schedule, we will try to make up as much time as we can, cutting our coffee break as short as we possibly can. But I do expect that we will run with our planned program a bit beyond the 5:30 scheduled time.

MR. CAREY:

Now that we have all been touched by the OMB sacrament and are in an appropriate state of grace, we can proceed. I wish I could have equal time.

THE AGENCY PERSPECTIVES PANEL (CONTINUED)

VIII. ECONOMIC DEVELOPMENT

THOMAS E. KELLY, Program Analyst,
Office of Program Evaluation, Department
of Commerce

MR. CAREY:

Now there are four speakers to be heard from. The day lengthens. The schedule becomes more flexible. But I am bound to get through this as well and as speedily as we can. We will continue the practice of sandwiching questions from the floor in as the speakers proceed. We will now have Tom Kelly, who is the designated hitter for the Department of Commerce. Bob Knisely could not be with us. I have seen the interviews in which Tom's comments were very, very lively indeed; and I expect more of the same this afternoon.

MR. KELLY:

Thank you. Sitting up here on the left hand of God, as it were, I got to look over Bill Carey's shoulder. I noticed that one of the notes his secretary made to him was that, judging from the interviews, virtually everybody on the speaking panel is rather long-winded; and he is going to have the time of his life trying to keep the time down.

Given that initiation, I will do what I can to be brief. I will resist what is an almost irresistible impulse to engage your natural fascination with the problems of evaluating tuna canning inspection and some of the other interesting things that we get to do at the Department of Commerce.

Before I get into my own remarks I want to clarify Sam Seeman's comment this morning about the Community Mental Health Center program, a program with which I once was associated. I want to make certain that everybody understands what Sam said, at least as I understood it. He said that the Community Mental Health Center program is one of the few, if not the only program that has been certified as an unqualified success by the Office of Management and Budget. I just want to note it so that no one leaves with the wrong idea that any of our good friends at OMB were looking for excuses to kill the program.

It will be a struggle to be extremely brief. I have a lot to say, I think, and it's a great enticement to take one's time talking to a group like this. But I am sure we will have a more lively meeting the more that you are involved and the less that we speak at you.

Bill Carey said this morning that there are a number of things that pass for evaluation. Oftentimes we get to talking about evaluation as if we all shared a common definition, when in fact we are dealing with our own personal or organizational conception of evaluation. The one understanding that seems to characterize all our thinking is that evaluation is a device which analyzes programs for the purpose of meaningful program change. It seems to me that of the many ways one can characterize and categorize the various activities that flow into program evaluation, there are two major streams. In the first place there is program evaluation research, whereby one tries to discover the objectives of a program, to determine what the resources are, to define the procedures by which those resources are applied, to measure outcomes from the application of those resources, and, when possible, to measure impact.

The second major stream in evaluation is any kind of analytical activity that develops facts about program design and performance for

the purpose of decision-making. It seems to me that in those two very rough definitions we find quite distinct characteristics. I think we all too readily assume that both of them are appropriate to the same situation. To the extent that evaluation is designed to promote meaningful program change, however, I believe the two types have quite different applications. I want now to reflect for a moment on the nature of program change in the Federal Government.

Bureaucratic change takes place for a lot of reasons. But two of the major reasons are these. First, some kind of shift in political philosophy sets in--a new person at the top, or a new set of policy recommendations, flowing not out of program performance as such, but from application of abstract principles in a way that dictates program change. I am not sure that program information gathered through evaluation is likely to be tremendously influential in that process.

The second way that program change comes about is through an historical accumulation of experience with the way a particular program runs. This is the argument concerning scientific change presented by Thomas Kuhn in a book called The Structure of Scientific Revolution. I am sure many of you are familiar with it. Kuhn presents a paradigm for the way in which scientific experience builds up and change takes place. His basic argument begins with an existing scientific theory. The theory explains a lot of the phenomena observed over time. As time goes on, anomalies creep into the observations. More and more things are observed which cannot be explained by the existing theory. People interested in a particular subject gradually become disquieted by what they find to be a less and less acceptable state of knowledge under the existing theory. Finally there is a breakthrough; a new

theory is derived that explains the anomalies and is therefore accepted by the field in place of the old. That is a scientific revolution in Kuhn's terms.

It seems to me that Federal programs follow somewhat the same pattern. But they are as much art as science. And because of this they derive at least as much of their energy and structure from social values as scientific theory. To my mind, Federal programs are essentially a patterning of resources and procedures based upon an assumed social value and a theory as to how that value might best be pursued.

Take the case of mental health, for example. If mental health services are considered a good thing, and we as a society decide that we need to invest in them, then an operative social value has been established. The choice of a particular configuration of resources, procedures, and objectives to pursue that value will be based, at least in part, on a theory of how best to define and deliver mental health services to appropriate recipients. Numerous constraints interfere with the realization of a theoretically pure delivery system, but compromises are made, and the program proceeds. Once the program is in place, the existing set of objectives, resources, and procedures becomes inextricably linked with the social value of mental health service. In the political arena, an attack on the delivery system is resisted as strongly as would be an attack on the social value itself. As in scientific revolution, major change is resisted until, in time, enough anomalies or inefficiencies are documented so that the method of service delivery is discredited without threatening the underlying social value. I submit that this paradigm fits the revolution in mental health service delivery which de-emphasized central hospitals and emphasized community services. Time and accumulated

information modified the environment for decision-making until a persuasive majority of the interested parties could agree that major program change was necessary.

Now, I think that the first kind of evaluation that I described, the rigorous type, is appropriate for developing the program history which contributes to the environment for program change. It seems to me that this is the essential function of program evaluation research as we read about it in many of the professional journals and as it is practiced as a specialty among many of the research corporations hired to do objective studies--not the least of which is MITRE. I don't believe that it's possible, in the complicated political and social environment in which we apply our skills, to construct a program evaluation, or even a series of program evaluations, which will provide meaningful, substantial, convincing information capable in itself of swaying a decision to change a major Federal program according to some prespecified decision date. This, to a lot of people, has been the expectation, the hope of evaluation. It certainly sounds like a logical expectation; but as we gain more experience with our Federal programs, I think we find that they are not so logically constructed as we assume; rather, they are patchwork applications of resources in the pursuit of social values. Research points up the anomalies, but only in the fullness of time will accumulated studies have their impact.

I was privileged to work with the Urban Institute a couple of years ago in attempting to find out exactly what the problems and the possibilities were in evaluating mental health programs. One of the things we found out was (and I'm using mental health simply as an example of other Federal programs) there was not in place the set of logical links between legislation, program objectives, resources, procedures, and intended outcomes that would allow a research design

to be quickly and successfully applied to those programs. I think that this is still true as I view other agencies. I am working currently in the Department of Commerce and I don't find there is anything particularly different where I am now. Program evaluation research is a tool, but it's a long-term tool. It contributes to a gradual accumulation of information about a program which may eventually result in a decision to change the program, but it will not do this in and of itself, and certainly not in the short term.

What do we have then? We still have a felt need to influence short-term decision-making in the Federal Government. Well, what is decision-making in the Federal Government? Is it a logical application of knowledge and principles to come out with the best possible solution to a knotty problem we all experience? We all are aware that decision-making in the Government is a political process, with a small "p" in some cases, or a large "P" in other cases. To that extent, it is a result of a conflict of interests which occurs in a chain--often a hierarchical chain made up of a certain group of people who are charged with responsibility over a given program, which may be fairly low in the bureaucratic hierarchy. These people, vertically aligned, take various positions relative to one another on any program decision in which they are all interested.

That position-taking or layering of divergent positions, is, I think, an important process. To the extent that it's a political process, to the extent that it's an attempt on the part of one participant in the decision chain to use knowledge to influence another part of that decision chain, it represents both a cooperative and an adversarial undertaking.

There was a question at today's luncheon gathering which I think illustrates the problem. The question indicated a certain lack of trust or acceptance of the statement that OMB is really interested in doing the right thing by programs. The questionner seemed to recognize that there are pressures on OMB budget examiners which are prejudicial to certain programs. There is no need to pick on OMB--one can find similar pressures at each level of the bureaucratic hierarchy. We each respond to the program manager for whom we work as staff. Our rewards tend to come from pursuing or moderating the interests, biases, and concerns of the manager for whom we work. Naturally, we do our best to base our actions on information which is as factual and objective as we can make it. On the other hand, we find that we are actually serving managers who are involved in a political process, who are attempting to influence one another, both above and below in the vertical decision chain.

Here is my major point, and I'll make it quickly. It seems to me that, if an evaluation office is set up to serve a particular manager and to satisfy the information needs of that manager about a program, and if that manager is engaged in an adversarial and cooperative process with managers above and below him or her in the line, then that evaluation office must provide information which is distinctly and specifically designed to meet the information needs and interests (in the double sense of that word) of that one specific manager. To the extent that the evaluation office is required to gather data and information on a short-term basis to affect a given decision, and to the extent that the information gathered is made available through some kind of a pseudo-line process to the evaluation staff office above, and above it, and above it, so that eventually it is common property--then that evaluation office has ceased to meet the specific interests and information needs of the manager for whom it works. It seems to me that if the information which the manager requests

becomes public information as soon as it is gathered, then it is probably going to be viewed by the manager as a threat to his or her autonomy--and be less useful to that extent. To the extent that the information is "intelligence," providing factual knowledge on a confidential basis, it allows that manager to be a much more effective position-taker.

I think that if an evaluation office is not set up to do long-term evaluation research and is nevertheless required to do formal, public studies to affect decision-making, it's likely to turn into an overhead function rather than a valuable, important part of the decision-making process. It is not in the manager's interest to provide an evaluation office a topic to study when the forthcoming information may be used against the interest of the manager that requested the study. As a result, the kinds of studies that the evaluator will be asked to do will be studies which are of marginal relevance to major program issues on which decisions are likely to be made. To the extent that topics for evaluation appear to be important superficially, there will usually be enough subtle communication between the manager and the evaluation office to establish that the nature of the study should not be such as to injure the interests of the manager.

I recognize that this theory smacks of cynicism. It needn't be applied cynically, however. The positive upshot of this analysis is to help us recognize and act on human factors which influence organizational receptivity to evaluation. All of us would be wise and fair in the absence of pressure. Under conditions of threat, however, instincts such as self-preservation often conflict with our more rationalistic leanings. Since managers are people, they react to pressure both rationally and irrationally--simultaneously. The organizational environment in which decisions are made is designed

to create stress and to enhance the competition for influence. Under such conditions, information--such as that gained in evaluation--may be viewed not only as a tool but as a weapon.

Here are the lessons which emerge from this reflection. To the extent that we construct hierarchical offices of evaluation, each higher office overlooking and using the products of the lower, we heighten the sense of threat which evaluation presents. To the extent that we conduct evaluation outside the context of "small p" political decision-making--as an objective program research and documentation activity, set apart from the management structure--we reduce the immediate threat and improve the prospects for long-term relevance. To the extent that we conduct evaluation within the manage-management structure as a low-key intelligence gathering effort for the use of individual managers, we are likely to improve its short-term relevance for decision-making.

I could go on, but I will end by reiterating that I think there is a role for "intelligence" as a definition of the information that we gather in evaluation, to the extent that we want to influence decisions. If we are content to influence decisions in the short term, it seems to me that we can often turn to a journalistic approach to evaluation--taking the example of a New Yorker profile which openly says: this is biased, this is personal, this is a one-shot view, but it does provide the information specifically required by this manager at this time for this decision. To the extent that we are trying to build a long-term program history, we will use something that is much more rigorous, much more scientific, which we call program evaluation research. That is really all I have to say right now.

MR. CAREY:

Well done. I guess I was wondering as I heard you talk whether the political people whom we cannot ignore view program evaluation as largely an ivory tower process. I think to the degree that that is true, it's a very heavy burden for evaluation to carry.

MR. STROMSDORFER:

If it's an ivory tower process, it's their fault because they don't interact appropriately with the evaluation. They won't specify program objectives. They won't specify program needs.

MR. CAREY:

You are including Congress and the Committee staffs and institutional offices of the Congress and all the rest with it?

MR. STROMSDORFER:

Pretty much. There is a major current of this. It isn't the only current, but it's a major current of behavior.

MR. CAREY:

I might take that point and that comment, but I also think that to the degree we over-theologize the whole business of evaluation, we contribute to making it spooky, unfathomable, tedious to read, complicated to understand. You know, you look at the life of a Congressman, you look at the life of even a Wilbur Cohen, 15 minutes is available somewhere in the day or the night to read something. The pretentiousness of a lot of the evaluation I have seen contributes to this ivory tower state of mind. I think we have to be very, very careful of it. Sometimes I have thought that while evaluation has an important role, an important place, policy change and even program change sometimes works just about as well when it comes out of an interactive, a very informal kind of a process. It's a process

of criticism. It's a process of response to criticism, of debate and argument. It is not as elegant by any means as what we are talking about as evaluation. It also has its place.

I remember one time we had been inventing The Great Society at a furious rate and whipping messages to the Hill at two-week intervals. The President had accumulated a whole truckload of those five-cent souvenir pens that we used at signing ceremonies. It was all a very exuberant time. We were flinging these programs out on state and local governments one after another. One day, I was visited in the Budget Bureau by six Directors of what we used to call the "PIGS"--the public interest groups. The Governors' Conference, Conference of Mayors, Council of State Governments--they call themselves the "PIGS" and they are proud of it.

PANEL MEMBER:

The corresponding group that you represent here is the "HOGS"--that is, high officials of Government.

MR. CAREY:

Thank you. I accept that.

We had a sedate discussion for a while about the problems of multijursidictional programs and multiagency programs. Finally, Bernie Hillenbrand lost his cool. (He represented the National Association of Counties.) He said, "Bill, if you really want to get this thing straightened out, why don't you have some kind of a policy rule in this administration that, as these great programs are being thought up, and as program changes are being thought up, that state and local and county people ought to have a voice in it and be consulted somewhere." I didn't have a very good answer. When the meeting broke up, I talked to a couple of LBJ's White House counselors.

They said, "Oh, we could never do it. The President wouldn't want to give away his options. He wouldn't want to telegraph them. He wants to have control. Don't even try it."

I heard them, but I wasn't convinced. I knew that my chief, Charley Schultze, was due to fly to the ranch the next day to have a working session with the President. So I had a word with Charley and gave him a draft of a short memorandum for the President to sign and send to the agency heads.

I said "You might take it up with him tomorrow if you get a minute." He said, "I'll take it with me." So he went off to the ranch. He was telling me later that it was a very, very hot day. The President insisted on giving Charley a personally conducted tour of the pastures, and the President was protected by very high boots. Charley just had his beat-up shoes on, and as he tried to sidestep the cattle droppings and keep up with the man, he was pretty well exhausted.

Then the President said, "Let's go to work." He gestured toward a picnic table alongside a clump of trees. It was a very, very hot day. The President pointed to the table, directly beneath the sun, and said "Sit down there, Charley."

So Charley sat down in the Texas heat with his pile of papers. The President climbed up into a hammock swung between a couple of trees. He is swinging in the hammock, and he's got his bottle of Dr. Pepper; and Charley is saying, "Mr. President, we've got this budget problem, and we've got that legislative problem," and he would hand up a paper to the President.

Finally, with the sweat streaming down his brow, he reached my little piece of paper. He said, "Now, Mr. President, if you'll take a look at this." He handed it up. The President began to read it as Charley said, "Let me give you some background on this." He got no farther. The President cut him short. "Charley," he said, "don't waste my time. Just hand me that pen."

I don't know what you think of that, but it's a little example, perhaps, of where you can accomplish something that does make sense, that does make a difference in the quality of management and administration without elegance or pontification of research and analysis; and I think there may be a place still for both things. Let's not, in glorifying evaluation--although I don't think we have done too much of that today--let's not rule out hunch and judgment where they can get the job done.

The next speaker is John Evans, who is Assistant Commissioner for Planning, Budgeting and Evaluation in the Office of Education. I think he has got something good to tell us too.

IX. EDUCATION

JOHN W. EVANS, Assistant Commissioner
for Planning, Budgeting & Evaluation,
Office of Education,
Department of Health, Education and Welfare

Thank you, Bill. I think your war story is very helpful, but I know it's not one I can top. Instead, I want to speak very quickly about the topic at hand, trying to use some history and a set of problems to speak to the question of what perspectives on evaluation exist in the Office of Education where I have responsibility for that function.

The brief history I want to recite should recall for all of you (and I think most of you don't need very much of that recollection) the principal fact that this gathering here today, this interest in evaluation, this surge in evaluation funds and contracts, this emergence of evaluation from fiscal, managerial, and programmatic obscurity to being something which is now all the rage, all reflect an historical change that has come about in a fairly short period of time.

I came to the Federal Government in 1961 when John Kennedy took office, and I have held a number of positions since then, most of which have related in one way or another to program evaluation in several different Federal agencies. It seems to me it's not an exaggeration to say that, as recently as a decade ago, the environment, the outlook, the attitude toward and the utilization of evaluation in Federal Government agencies on social action programs was entirely different than it is today. I see some of my old OEO[15] colleagues here, and we can certainly hold old home week on that score.

[15] Office of Economic Opportunity.

Perhaps the best way to make this point is to slightly caricaturize the change that has occurred. I might try to sum up in a single hypothetical example, a caricatural one to be sure, what the situation was like as recently as a decade ago. Those of you who have been in the process, I think, can probably support what I am about to say.

If you go back ten to fifteen years, what you would find is a situation pulling all the problems and evils together which is something like this. You have a Federal agency in which the head of the agency decides, either reluctantly or willingly, that an evaluation needs to be done on one of his programs. He summons one of his top people and says that either OMB has told us it wants, or the Congress has told us that they want, or I personally want, an evaluation of this program.

The first thing to note (as others have observed) is that usually no agency evaluation mechanism of any consequence exists to which he can address that question or that task. If one does exist it is buried somewhere in the bowels of the organization. Finally, somebody says, we'll do it, and the task is entrusted to someone who is a program director or administrator. Finally, an RFP is issued. However lengthy and wordy the RFP may be, it says really little more than, "please submit proposals to evaluate this program." In response to that kind of lack of specification, in come a series of proposals from academic research institutes, commercial research organizations, and the like, which range all the way from $25,000 to $2.5 million, and all the way from quick and dirty site visits to sophisticated, experimental-design, longitudinal studies.

How those things can be compared and one chosen among them is hard to imagine, but that task gets done. One is chosen. The contract is signed, and work gets underway.

After that the thing is generally lost from view since there is no one to oversee it or direct it, and it has no organizational home or responsibility. Some substantial time later, in comes a report. The important thing as far as the evaluation process is concerned is that the report is too late to influence the decisions which gave rise to the need for the evaluation in the first place; it is too voluminous to be read by anyone who would be in a position to make those decisions; it's too technically esoteric to be understood by them if it were on time and they were to read it; and there's a good chance it has become irrelevant policywise to the issues which triggered it at the outset.

The results are that, first, it goes on the shelf where it is unused and uninfluential in policy, program, and budget decisions. And second, even worse, when its existence is belatedly and critically recognized, it contributes negatively to the reputation of evaluation as useless.

That, as I said, is a somewhat caricatured example, but it summarizes the set of problems that evaluation in the past has had, and to some extent still has, to deal with.

I can sum those up by saying that, first of all, there is the problem of resources. For evaluation to be effective, there must be adequate fiscal and personnel resources at the agency (or at whatever level) for it to be carried out. I will come back to that in a second.

The second major problem is that evaluation must, as we have already discussed earlier this morning, be situated in an organizational location where it is possible for two things to occur: (1) objective and technically competent evaluations can be conceived and carried out; and (2) there is an avenue of influence for their results to impact budget and policy decisions. Therefore, evaluation, in my judgment, clearly has to be one of the principal executive staff or decision-making functions--the other being planning, budgeting and legislation--which must be lodged in a position where it can have that kind of access.

It's worth digressing here to say that even when all those conditions are satisfied, evaluation findings and activities will get nowhere if the head of the agency in question is not himself or herself personally interested in making use of those findings for managerial and decision-making purposes. That I think is still another thing that has changed substantially over recent years.

A third problem that must be dealt with is the matter of competent methodology. Evaluation is a term that means many things to many people. Evaluators, like ladies of the evening, suffer a great deal from amateur competition. What has to happen is that the function cannot simply be some casual kind of activity. When we talk about program effectiveness, we are basically talking about a cause-effect question. We want to measure what changes have resulted in connection with the program, but more importantly we want to be able to attribute those to the program, not just the passage of time or some other extraneous variable. That immediately brings you into the matter of research and evaluation design.

The other reason why design and methodology are so important is because all of the programs that we are talking about (or nearly all of them) are inherently controversial social action type programs. As such, in the political sphere, in the Congress, and in the public, they have both their protagonists and their detractors. That means that any evaluation of any of these programs, no matter what it finds--whether it finds the program effective or ineffective--is going to be attacked, not because the findings are distasteful which may be the real reason, but on methodological grounds. Therefore, if the evaluation is not itself methodologically defensible to a reasonable degree, its chances of influencing policies and budgets is thereby lessened substantially.

Fourthly and finally, there is the problem of dissemination and utilization. Even if you are lucky enough and smart enough to do everything right from beginning to end in terms of resources, personnel, design, avenues of influence and so on, it's not automatic from there on at all. The inertia in Congress is tremendous. The mere production and dissemination of findings, however intellectually or methodologically compelling they may be is usually not enough to sway a decision, change a program, alter a budget, or change a law. There must be other kinds of mechanisms to affect that.

Moving along very quickly, then, given the basic history of evaluation as I have personally seen it, given also the central problems that surround its implementation and use in Federal programs, what we have tried to do at the Office of Education is develop a mechanism to deal with or minimize those difficulties and problems.

What that means is that, first of all, in the matter of resources, as far as we and many others today are concerned, because of the historical changes which have occurred, many of us can no longer complain about the matter of resources. It is true that in the Office of Education, we don't have all we need. We have maybe 25 or 30 people that can be called full-time professionals allocated to the evaluation functions, people with advanced degrees in the behavioral sciences, quantitative analysis, measurement, sampling, and the like. We have an annual budget, coming from a separate planning and evaluation appropriation, plus set-asides from program funds, which comes to about $15 million. But this must be used to evaluate an $8 billion budget which embraces over a hundred programs.

While resources are not luxurious, contrasted with the situation eight, nine, ten, fifteen years ago, we cannot really say that the principal obstacle to accomplishing useful evaluations is a lack of resources, though certainly it remains a problem.

On the matter of organizational location, the evaluation function is in the Office of Education coupled with those other functions[16] that I mentioned earlier. I am the Assistant Commissioner for Planning, Budgeting and Evaluation. I also occupy another position on an acting basis which oversees the Office of Legislation. All of those functions are combined together, and I report directly to the Commissioner of Education. So once again, at least in our case, that cannot be used as an excuse for why evaluation isn't progressing or doesn't have the opportunity for influence. I mention these because it is my impression that these ways of dealing with the problems I have mentioned are far from universal in Federal agencies at this time.

[16] Planning, budgeting and legislation, see page 112 above.

On the matter of competent methodology, again the key in my judgment is to assemble the kind of technically qualified staff I have described, and then to develop a system which consists of people like that designing the evaluation in-house. That is, we design it down to specifying such things as sample size, control group procedures, and types of outcome measures. That kind of highly descriptive and prescriptive detail then goes into an RFP which is issued for the field work, because obviously very few Federal agencies can function like the Census Bureau. The work is then carried out under contract through the competitive procurement process.

Finally, in the matter of dissemination and utilization, we have developed a system where the person who is responsible for designing the evaluation in the first place chairs a technical committee to review the proposals which come in on it, is responsible for very close, hands-on technical monitoring of the instrument development, field work, and analysis while it is going on, and is then finally responsible at the end for writing a layman-level summary of the results as they come in from the contractor. I think it's a mistake to try to use contractor reports as the principal vehicle for disseminating or communicating evaluation findings. We write brief, layman-like kinds of summaries that are then sent to all members of all four Congressional Committees which oversee our programs (both Authorizing Committees and Appropriations Committees), as well as communicated widely within the Office of Education, HEW, OMB, the Domestic Council, and the like.

Even that usually won't do it. We are now experiementing with a further effort to get evaluation results to actually affect decisions and budgets and program guidelines. It's a small and

essentially bureaucratic device, one we call the Program Implications Memorandum, or PIM. What we do in addition to the summary is write a memo which extracts what in our view are the program, policy, legislative, and budgetary implications of an evaluation. It's an action memorandum, signed by the Commissioner, which in effect says, "All right, the evaluation findings indicated so and so. That means we should prepare a legislative modification. The Office of Legislation will be responsible for doing this by November 30th. The budget should be changed in the following way. The regulations should be changed in the following way; these tasks are assigned to these offices and they must be completed by such and such a time," and so on.

We have yet to really develop this mechanism, but I think it is a promising effort to overcome what is, as I said before, a major problem. Even once you have got timely, methodologically sound, and policy-relevant findings, they won't implement themselves.

I just want to close very quickly with a couple of other remarks that have been prompted by some of our discussion so far this morning and at lunch. I am sorry Toney Head didn't stay and we didn't have more of a chance to talk with him and question him about OMB's role, because one of the very serious problems bound up in the dissemination and utilization problem mentioned before is that of credibility. We had an incident, I remember, not long ago when President Nixon was forwarding one of his budget messages to the Congress. As you all know, there have for the past few years been proposals by the current Administration to reduce expenditures in a number of domestic programs including education. The thrust of the budget message to the Appropriations Committees and to the Congress was: we are proposing that certain of these programs that are overseen by the Office of Education either be eliminated or reduced because they have been found to be ineffective.

That message was composed and sent forward without the benefit of counsel from us. So back from the Congress came a formal request to the Administration, OMB, and the Secretary of HEW which said, in effect: that is very interesting; would you please send us the evidence and materials that cause you to make the judgment that these programs are ineffective and therefore candidates for elimination from the current budget?

We were then asked by OMB to produce such data and information, and we replied that there were no such data. Indeed, some of the programs in question had contrary evidence that indicated their effectiveness rather than their ineffectiveness.

Let me finish the example. It goes on. What happened was that we were unable and unwilling to produce the nonexistent negative evaluation data, and so certain things were concocted by others and sent forward in response to Congress. They so offended the Congress in their patent irrelevance to the matter of effectiveness and their unpersuasiveness as objective and empirical evaluations that, in effect, the Congress said, if this is evaluation, we'll take vanilla.

This, in turn, led many people in Congress to the opinion we were talking about earlier, which is that things called evaluation submitted by an administration or submitted by an agency are inherently untrustworthy. During that fiscal year, we received a substantial cut in our evaluation appropriation which I think can be attributed largely to the set of events I have described even though we had been in an historical trend of increasing evaluation appropriations and attention.

MR. CAREY:

Better there than in the programs.

MR. EVANS:

Well, possibly. So the matter of credibility is extremely important.

I just want to finish up with one final observation, and that is to add my view to a couple of points that have been made so far on how evaluation fits into major decisions in the Federal Government and what its outlook is. I think the views that a number of speakers have expressed so far are quite correct in emphasizing the fact that decisions on these programs, on their supporting laws, and on their budgets are inevitably and inherently a political decision. We function in a pluralistic system in which the findings from an evaluation, even if they meet all the good criteria that I have talked about, still are, and I suspect always will be and should be, only one input into a decision which is a pluralistic and political one. And those of you who are freshly getting into this field or haven't been in it long, if you become easily disillusioned or are naive in thinking that evaluation findings constitute an automatic decision-making mechanism, I think you should disabuse yourselves of that notion. On the other hand, I don't think that the fact that many and perhaps even most decisions will be predominantly political, rather than pristinely rational based on evaluation findings, should lead us to excessive cynicism that evaluation is not worthwhile or cannot be effective. There are long-term trends in society, in the Government, and in the Congress (for example, the introduction of the new budget committees in Congress), all of which indicate that there is a movement toward the rationalization of decision-making, policymaking, and resource allocation; and that while evaluation findings will not always be used fully, and sometimes not at all, they will be used more and more. They are needed more and more; and I think those of us who are in the business of providing them will, while we may lose a lot of battles, stand a chance of winning some too.

Let me stop there and try to answer some questions.

X. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

MR. CAREY:

Let's take a few questions to Tom Kelly and John Evans.

PARTICIPANT:

Walt Bergman, IRS. The question is to Tom Kelly. Your prescription for serving your master, if you will (and I'm paraphrasing in terms useful from his vantage point), how do you square away that point of view with the operating-in-the-sunshine and freedom-of-information kind of environment in which we live today? It may be that I misread you, but it seems you are saying information should not be used against the official to whom you are directly reporting, the bureau head or the agency head. It would seem to me that the best way you can serve your master is with objective evaluation because it is going to be used by others and it's going to have to stand the light of day.

MR. KELLY:

I wouldn't want to suggest that it is impossible or completely infeasible to do objective evaluation at a given level of the bureaucracy in good faith. What I would suggest is, to the extent that a decision is a hard-fought decision, it is likely that the evaluator will be under pressure to fuzz the question, or study a sub-issue in place of the central issue at hand, the more general the distribution of the knowledge is going to be.

What I am really arguing is that if one expects that evaluation is designed principally or even solely to affect decision-making and to lead to program change, one must consider changing his style

of operating. It seems to me that the kind of evaluation research which John Evans described as being done at the Office of Education is an extremely valuable kind of social research which will have occasional short-run advantages, but more likely will have impact in the long run. John is certainly willing to correct me. I haven't made a study of the impact of his work. But I would think that it should be fairly judged on its long-term, not its short-term impact. If the interest of the evaluation officer is specifically to affect a decision by changing a position taken by a relevant somebody, he must recognize that there must be a mixture of the public information with not-so-public information so that the individual who is taking a position in the decision-making process will have a slight competitive edge. There always is, it seems to me, a certain pressure on a staff person to help the boss make good decisions without foreclosing future options. It's awkward to call it serving one's master. On the other hand, to the extent that the evaluator tries to be totally objective, without regard to the interest of his master, evaluation becomes something to be tolerated and thrown a bone. Viewing evaluation as an overhead item, the pragmatic manager merely requests studies which will not hurt in the short run and could conceivably help in the long run.

PARTICIPANT:

Mark Markley from Stanford Research Institute, also for Tom Kelly. I was interested in two things you said or talked about touching on the Urban Institute study at NIMH[17] and the other one talking about Tom Kuhn's theories and paradigm change.[18] Could you comment briefly on what you see the impact of the Urban Institute study being, specifically in terms of any changes it may have introduced into the Zeitgeist in Washington?

---

[17] See page 100 above.

[18] See page 98 above.

MR. KELLY:

It is hard to know precisely what the effect of any given study is. I find it difficult to discriminate changes in the Zeitgeist from changes in my own personal world view. I can only comment on what I learned out of taking part in that study and on the kinds of attitudes which I have encountered in subsequent conversations with people engaged in evaluation.

When I first participated in that study, I had been in the Federal Government for approximately three to four years. I had been a management intern. I had worked in personnel. But I was still strongly convinced that Federal programs were a very logical kind of thing. That you looked at the objectives and you looked at the resources and you applied the resources in certain ways. You were going to change things that were measurable. It's unfair to pick out the Mental Health programs because I think they are typical and shouldn't be singled out for this quality. What we did find in analyzing those programs was that the logic simply wasn't complete--that it wasn't necessarily provable or demonstrable that the application of resources in a certain way was going to "improve the mental health of the American people," for example. That is an unmeasurable objective, so the procedures were not shown to be particularly well chosen to achieve that goal. How can you achieve a goal if you don't know what it is once you have gotten there? We have a lot of trouble in even defining mental health and mental illness except on an individual basis. An individual might be variously defined as mentally ill or mentally healthy depending on whose standards you apply. But you don't talk about national standards for defining mental illness, at least not in a democratic society.

What that study did for my own personal world view was that it complicated it a great deal. It led me to wonder whether it was

possible to use the methods, the principles of the experimental method, to apply in most cases to interventions in the social process. I woke up a little bit. I lowered by expectations of what evaluation could do. Perhaps they are a little bit too low at this point. I admire what John Evans describes his staff as doing. I think in the long run, that is probably the only way that we are going to have a strong contribution to national growth in terms of our knowledge and our theory of social program intervention. In the short run if we are really interested in saying that we have a role in a bureaucracy, and that bureaucracy, in the short run, is designed to resolve conflict and make decisions, I think we have to lower our expectations of the art and be willing to scrounge around a little bit and say, "What I am giving you is biased, what I am giving you is personal; but what I am giving you is eyewitness; and it is the best dope that I can give you right now on how that program is performing in the field. Use it as you will."

PARTICIPANT:

I am Bob Crain with the Rand Corporation. While I am a very strong believer in the notion of an insulated evaluation group much like OPBE, I wondered if Mr. Kelly's concern could be met by having as one of the objectives of the evaluation not only an accurate and high quality report, but also building into the evaluation process more of the kind of face-to-face or personal contact which Kurt Lewin would say is necessary to help a program manager accept the recommendations. I wonder if you'd comment on that?

MR. EVANS:

That's a hard one to comment on, Bob. I guess all I can say is that, as you have sensed and as Tom has properly inferred, we

take as our guiding basic model of evaluation the one you were involved in carrying out for us, namely, some version of the classic model of experimental design, where even in the natural setting, if possible, you can achieve the condition of random assignment to treatment and control groups and thus eliminate one of the serious haunting problems of all evaluations which is ambiguity about the estimate of the effects of non-treatment, or the noncomparability of control groups. I argue with Don Campbell about that on occasion. He thinks this is generally feasible, but I often take a version of Tom's view which says in effect that Federal evaluators have got to realize first of all that in the Federal Government, decisions are going to be made either in the presence or absence of information. Therefore, the evaluator's task is not necessarily that of conducting the perfect evaluation. It is rather the task of information getting to the decision-making point which will improve the decision. Sometimes that may have to take the form of a one-day, quick and dirty site visit. That, of course, is a serious retreat from what you'd like to do. It's fraught with ambiguities, and so on. So my sense is that you do that as best you can, which, of course, is an experimental or quasi-experimental study whenever possible.

As far as the other levels you are talking about, you raised another question which is what levels of decision-making a single evaluation can properly serve. Again, I would have to say that our evaluations have been primarily oriented to what I would call Federal decision-making issues. They are oriented toward the Congress and the Executive Branch. Is this program working as a national effort? Should it be expanded or contracted? Should it be eliminated? Should it be reformed or changed?

Now, people at the local or project level have different issues and needs. One of the raging questions that will always be true in evaluation is, can you conduct a single evaluation which embraces the information needs of these different administrative levels? I am inclined to think that is rather hard to do, though sometimes it is possible; and wherever it can be done, it should be done.

MR. CAREY:

A brief question from Bob Hemmes and a brief response. That will have to end these questions.

MR. HEMMES:

Apropos of the remarks, Mr. Chairman, made by the panel and the participants regarding objectivity and who is going to do the evaluation, I'd like to call your attention to William Sarcefield's article in the March, '76, "Government Executive" in which he asks who is going to do the evaluation. He said "if you do it in-house, you can't evaluate your own boss. If you do it at a university, faculty members are not good prospects for applied research tasks. They tend to turn the task into basic research in line with their own interests; and if you go to a consultant, the consulting firm is likely to be oversensitive to the decision-maker's wishes. Instances have been observed where a consulting organization asked to evaluate a program provides its client with a white-wash which the evaluator assumes the client expects." (The latter doesn't apply to the MITRE Corporation of course since they are in the honesty business.)

MR. CAREY:

I will rule that that was not a question. That was a contribution.

XI. MENTAL HEALTH

JAMES W. STOCKDILL, Director,
Office of Program Development and Analysis,
National Institute of Mental Health

MR. CAREY:

We are going on now, and we are going to hear from Jim Stockdill who is the Director of Program Development and Analysis at the National Institute of Mental Health. Jim, if you will proceed without any further ceremony, I'll be grateful.

MR. STOCKDILL:

Coming eighth in the batting order and coming at this time of the day, I am not sure I can keep you awake and interested. In fact knowing what I have to say and how I am going to say it, I may not stay awake myself. I'll do my best and try to talk loud.

I have been involved in one way or another with the formal evaluation program at the National Institute of Mental Health since about 1968 when we were first authorized to earmark one percent of our Community Mental Health Center grant funds for evaluation. This authorization was done through legislation; so it was Congress that really got us started in the evaluation that has been referred to two or three times today, and, which resulted in an OMB finding that we had a very successful program.

Since 1968, we have changed our philosophy several times about how to use what we call one percent evaluation funds. We have varied both the process and substance of evaluation. During this period we have supported about 80 to 85 distinct projects costing around $7 million. You can see that we are not in the business of supporting

large studies, but small projects. The only thing all these projects had in common was that they were all supported by one percent evaluation money. I am not trying to make it sound like we have a nice concrete evaluation program.

In fact, I've come to the point where I really dislike the term evaluation. I think I began to dislike it three or four years ago when it became popular to evaluate evaluation. We did it. Several other Federal agencies did it. We funded the Urban Institute study at NIMH which was referred to today, and another one as well. We got some free reviews from universities. This was proceeding all right until the Director of the National Institute of Mental Health asked us to evaluate the evaluation of the evaluation. We did that, and gave him a report, and thought we had done a good job. But he said, "It's only fair to send <u>that</u> evaluation out--" You can see how it kept on going.

I am not trying to discourage evaluation of evaluation, but let me advise you that if you are starting a new program or getting into launching an evaluation office, once they come to evaluate your evaluation, don't fight that; but don't get caught up in trying to evaluate the evaluation of the evaluation.

Now that I have that off my chest, maybe we can get down to real business here today--which I think is--what have we learned from our experience with these evaluation funds? It seemed to me when I thought about this meeting, that when I have participated in other sessions like this in the last two or three years (we have had a conference in each of the HEW regions on evaluation), what we always talk about is the problems that we have had, the problems in methodology, the problems in timing and setting priorities and so on. So today I'd like to try to do something different and see if I can

identify a few positive things that we have learned about how to use evaluation. I will try not to repeat too much of what has already been said.

Maybe I'll just say it in a little different way. It seems to me the first thing we have learned about evaluation and decision-making as a process is that managers and evaluators must continuously keep in mind the political origins of the program being evaluated and not just look at what they decide are the current goals and objectives of the program at the time of the evaluation.

We had a very difficult time on this with the Community Mental Health Centers program when we first started to try to evaluate the program three or four years after it came into operation. We had to go back and reconstruct the objectives of the original legislation. Over the years many of the original concepts and objectives had grown fuzzy. I think the evaluator who doesn't do his or her homework on the historical and political development of a program is largely going to find himself ineffective in designing useful evaluation activities.

Also important for any specific program evaluation, or piece of a program that is being evaluated, is to clearly identify which decision-makers you are trying to influence. The evaluator must know who really has the power to make a change or actually decide to continue or discontinue a certain approach.

Let me again use the Community Mental Health Centers program as an example. Over the past few years, the approach that we have tried to develop for evaluating those centers has been directed at two major levels in the total hierarchy. The first group is the Congress. It is Congress which decided in the last few years whether there would

be a Federally-funded program, and Congress which decides how much is going to be spent on that program. It also decides whether there will be any basic changes in the authorizing legislation.

The second group we tried to direct our evaluation studies to was the managers or directors of the community centers themselves because they are the important level that determines how resources actually get allocated.

We realized, at least during the last few years, that the Federal and state bureaucratic levels in between the Congress and the local levels weren't really having that much influence in determining how the program was being operated. So we tried to develop studies that would hopefully help the two levels mentioned above to make wiser decisions on how the program should be designed and how it should be operated.

I think we have also learned that evaluation activities in a bureaucracy can be a constructive source of conflict. By that I mean that they can serve to smoke out what have been largely hidden conflicting ideologies and interests, particularly in programs that have been established for quite some time. By introducing some kind of a systematic quantitative analysis and discussing these with the program managers, I think you can surface a lot of these problems. Once you have done this, the ideology of a program and the support of interest groups will no longer really suffice by itself to maintain an ineffective program. Even though the program may survive, you have a better chance of reducing the level of resources that might be devoted to the program than if the evaluation had not been done at all.

We have seen this in the National Institute of Mental Health in relation to some of our training grant programs which have been in existence since 1950, or the late 1940's. Some of the traditional, older programs are really no longer justified (at the same resource levels) in any objective way that you can identify. However, the staff operating these programs, as happens in many places, have become over-identified with the program itself, with the universities that were receiving the training grants or with other constituency groups that had grown dependent on the program. The original purpose of the program has been lost in this whole long history, but in cases like this, by surfacing some of the conflicting views through an evaluation and planning process, I think you can sharpen the judgments and improve program decisions.

As has been referred to today by several different speakers, including the OMB luncheon speaker, a frequent obstacle to effective evaluation is the fact that the program's objectives and purpose just weren't clearly defined in the legislation. The general response is to just curse the fuzzy-minded politicans or administrators that started the whole process and then the evaluator may go off in a room and write down his own objectives for the program--just to satisfy the evaluation process. This is called the "phantom" approach to evaluation which often produces interesting but not very useful results.

We have found that a useful approach that the evaluator can take, and usually there isn't anyone else to do it, is to try to go back and reconstruct and create a new picture of all the human needs, political and social interests, and theories that formed the basis of the original authorization of the program. Developing a description of the inputs, of the program activities and of their desired relationship to program outputs or social change is a useful role for

the evaluator to play. A systems approach to a kind of meticulous specification and redescription of the inputs (who participated in developing the original authorization, how do the current conditions differ, etc.), can be a very useful function of evaluation. It can at least be very valuable in helping direct future program legislation in the same area. It is a very frustrating and time-consuming approach, but I think it will cause less conflict in the long run than what we called "the phantom approach," where the evaluator sits down and develops objectives that fit his evaluation process.

Early in our evaluation experience, we romanticized evaluation as an objective scientific process. We felt it should be uncontaminated by political compromise and based on some kind of an intellectual power. We found that the intended effect of evaluation programs or the effect of evaluators, if you are going to determine effectiveness by influence on policies and decisions of administrators, is seldom if ever totally objective. The evaluator is either trying to find a weakness in a program, trying to justify a program or, in a lot of cases, trying to further the field of evaluation itself and his or her role in that field. I think there is a quote from James Schlesinger, formerly of the Defense Department (I have never had anything to do with the Defense Department), which supports this experience of ours. He indicated that, "In understanding the results of evaluation, we must bear in mind that analytical work is performed and decisions are reached not by disinterested machines but by individuals with specific views, commitments and ambitions."

The point is that the administrator must assume that the evaluator is something less than totally objective. In fact, I think if the evaluator doesn't care one way or the other, he probably wouldn't do a very good job of ever assuring either that a study got carried out (if he is doing the study type or survey), or that the results were brought into the decision-making process.

There is a political scientist from England who has been in this country the last few months evaluating our elective process and evaluating the role of the press in covering the election. He has indicated that there must be some commitment that drives curiosity and perception or there just would not be good coverage by the press. I think one can say the same thing about the evaluator. There has to be some kind of commitment there that drives his or her curiosity. It is something much different than scientific objectivity.

We have had evaluations of some of our programs by Mr. Nader, by GAO, and by the staff of the House Appropriations Committee. They have all looked at the Community Mental Health Center Program, some of them more than once; and there are our own studies of the same program. I would say none of them is objective. But if you put all of these reports together, they are all speaking from a different motivation, a different perspective, I think we can learn a great deal. The problem is having the time to pull all of these studies together and synthesize the results, if you will, to see what can be learned from them. We are usually in the bind of doing some studies, starting up others, and we don't put enough emphasis on analyzing all of the different findings and recommendations together.

Let me say a couple of things about utilization of results even though this has already come up several times today. Let me emphasize we have had many studies that have had no utility whatsoever. But I said at the beginning, I want to emphasize the positive side rather than the negative.

We have completed projects that have been useful as inputs into the development of new program regulations, and as inputs for changes in legislation which authorizes community mental health center service to children and the aged. There were studies that contributed to

changes in legislation and development of regulations in those specific areas. But I can't think of any case where we have done anything that would answer comprehensive questions that would make a change in a total program. What has been useful are projects which were directed at carefully delineated questions about discrete program areas or functions. That is the kind of study that has yielded useful information. There is no way, I think, that we could currently design a study to answer the comprehensive question: are Community Mental Health Centers generally assisting the communities they are located in? That is too long-range a proposition, there are too many uncertainties.

Let me leave you with the following summary of thoughts. Evaluating any social or human service program is primarily a planning or management activity, only secondarily a scientific activity. Evaluation should be a conscientious systematic effort to inform administrative and political decisions. I don't think we should think about it as research to improve some general level of knowledge. I think to lose sight of that reality will result in increasing amounts of information about interesting but unimportant questions.

I'd like to reemphasize what several people have already emphasized--that evaluation is just one input into the decision-making process; but, it can be useful in sharpening the judgment of the decision-maker. However, to insure its utility, the evaluator needs a lot more than technical evaluation skills. He has to understand the bureaucratic organization and the political processes within and without that organization. There can be a plurality of different kinds and levels of evaluation. If they are all pulled together somehow, it can be very useful in terms of incremental decision-making.

I would just like to comment on one other thing. Someone raised the question this morning of evaluation of research.[19] We have had some experience with that, but not much success. A very simple approach that seems to be effective has been to pull together groups of researchers around a specific area or issue and let them deal with each other about why they are doing what they are doing, and where they think the field should go. We did that this past year around the area of mental health problems during early or preadolescence (ages 9, 10 and 11). A lot of people were concerned that there are more and more emotional problems showing up in that age group but little research going on. We identified what we felt were 18 or so researchers around the country who were doing some work related to that age group. We brought them together, let them talk together in a conference like this, only of course much smaller, for two or three days; and we did get some useful analysis for new program direction out of it. The participants also felt that they improved their own individual research projects by having to bang heads with their competitors. A simple, but, I think, an effective kind of approach. Thank you.

MR. CAREY:

Thank you, Jim. That was a very balanced story. And that is helpful because I think what we have been trying to achieve here is just that: a balance.

---

[19] See Mr. Weinhold's comments, page 33.

XII. JOB TRAINING AND EMPLOYMENT

ERNST W. STROMSDORFER,
Deputy Assistant Secretary for Research and Evaluation,
Department of Labor

Mr. CAREY:

We are coming down now to the end of this session, and I suppose the idea of today was to get you all sharpened up for the workshops to follow. I certainly don't want you to go into those workshops in a state of alarm and despondency. That leaves it all up to our clean-up hitter, Ernie Stromsdorfer, who is Deputy Assistant Secretary for Research and Evaluation over in the Labor Department. He has a very simple task from the Chair, and that is to create an atmosphere of elation as we bring this afternoon's panel to an end.

MR. STROMSDORFER:

As with you, Bill, the kind of evaluation I am talking about and would like to get going in the Department of Labor, and perhaps in Government as a whole, is an interactive process among policy-makers, program managers and the providers of information. When I am talking about evaluation, I am not talking about the nuts and bolts of running a particular research project or experimental design project. I am talking about the process of providing information to aid in social decision-making.

What are the ingredients of decision-making? Information is one ingredient, and the political pressures that surround a situation are the other ingredients, if I can abstract a little bit. Basically,

the political pressures come from vested interests who claim that a given program will aid them a lot and harm others only a little bit or maybe not at all.

But what do all programs do if they are significant programs? Regardless of their institutional or programmatic structure, they do one major thing. That is, they redistribute income, and they redistribute social and political power. In the process of redistributing income and power, they also affect the structure of production, economic efficiency, the level of economic activity and a host of other social institutions--social institutions in a Veblenist sense. Patterns of behavior, patterns of conduct, ways of doing things, both social and psychological, and what have you.

In a context such as this, where the enlightened self-interest and the altruistic rapacity of vested interests attempt to influence social policies, the role of information, as I see it, is to make sure that self-interest remains enlightened and that rapacity continues to be tempered by altruism. It's understandable then that evaluation, or rather more broadly, the provision of information, is a highly politicized process. There is nothing necessarily negative about this thing. It's just a statement of what I perceive, and I am sure it's not a very startling statement.

Evaluation and the provision of information occupy a very ambivalent love-hate position in the Government. It suffers from the hypocrisy of a positive social ideology derived from the Enlightenment and other philosophical strands, coupled with underfunding and often misdirected funding. (I had previously written in here," consciously misdirected funding," but I guess it's not necessarily conscious. It just happens through the interaction of various groups.)

The methods of shortcircuiting the provision of information that might reveal the true effect of activity are legion; and when I approach the problem of dealing with evaluation at my agency, my fundamental operating principles are the following: I assume that program managers have a taste for uncertainty. They tend to prefer the uncertainty in which they remain essentially unaware of their ignorance, of what they don't know, to conscious awareness of what they don't know. There are thus two kinds of uncertainties.

The first kind of uncertainty does not necessarily restrain a person in decision-making or in pursuing his or her interests. Whereas the latter kind of uncertainty, informed uncertainty, tempers decision-making and probably constrains behavior somewhat.

I guess secondly, an operational principal is that bureaucrats (including myself) prefer a quiet life; and one of the ways in which they tend to insure that they have a quiet life is by arguing that political problems of one kind or another constrain activities, and therefore you have to go slow. You have to be careful. You have to consult with everyone and touch all bases.

Finally, I operate on the principle that it is <u>not</u> ignorance or basic incompetence which keeps us from getting the required information to aid in decision-making. Though it is true that resource constraints do pose various problems because most of our social programs are multi-dimensional, have multiple impacts and often the data base, the informational base which you need to find out what is going on, implies the absorption of the Gross National Product to achieve it.

I have a basically negative view as to the efficacy of evaluation and of the long-run prospects for providing sound information to the decision-making process. Let me give you some examples of what I mean.

My examples are necessarily drawn from my immediate experience in the Department of Labor. We have a regulatory program in the Department of Labor which is designed to improve the health and safety of workers in the society at large. There is a clear-cut social problem here. There is a clear-cut role for Government here because of the potentially enormous social cost that can accrue to individuals in society as a result of third-party actions.

Yet we see in the operation of this program what appears to be a stalemate due to the social, political and economic conflicts that have arisen among those who will gain from the program and among those who stand to lose from the implementation of this particular social program. There is a serious social conflict here. It is possible, although not absolutely certain, that more information on the economic and non-economic costs and benefits of administering this program might tend to reduce the level of conflict and make the course of action with respect to this program more clear. Apart from gaining an understanding of what is happening, the reduction of conflict and elimination of the stalemate itself would be salutory for the democratic process. But here is where problems begin.

In this program and in other regulatory programs in the Department of Labor, the nature of what one is attempting to achieve is not well understood. This lack of understanding begins with the very initiating legislation, as I believe Mr. Hemmes pointed out just before lunch. Congress passes laws which are very non-specific, and then the bureaucrats and the administrators proceed to the making of the real laws.

In the process of making these real laws, they have little guidance from the legislative history because within the legislative history, priorities are unstated. It is true that issues are raised and discussed; but priorities are unstated. So the people who write the Federal regulations have little guidance in their writing of the law and expanding of the law.

Well, then, a successful program manager, one who wants to get information about how to operate and manage his program, has to know what the intended and likely unintended effects of a program are. What data can be generated to describe these? Well, this question, as it is posed for the Occupational Safety and Health Administration in my judgment is basically unresolved after about six years of program operation. Reading the legislative history will not give you much enlightenment as to what we ought to do here since the debates do not assign relative priorities to the issues discussed therein. They don't lay out the former structure of the program either. That is one problem.

The other problem is understanding the process whereby the program is intended to achieve its effects. What data are necessary to describe this process? What is the program delivery system and how does it operate in society to achieve its end?

It is in answering these two above sets of questions that all evaluations and all searches for information on which to make a decision, whether rational or not, break down. And here is where the Government at every level and branch has the greatest potential to facilitate or shortcircuit the effort to gain information on how a program is operated.

We in the Office of the Assistant Secretary continually struggle to get program managers, data system developers and agency evaluation shops to ask this basic question set. We are not uniformly successful. Most of the program data sets as a result are fundamentally inadequate to understand programs. They are fundamentally inadequate as a base upon which to set up the more classical program evaluations. We cannot even well describe the structure and integration of program inputs, much less describe what the final impacts of programs are.

I want to stress again that the ultimate failure of most evaluations is a function of the failure to develop adequate program process data and to adequately understand the program process. I could go on and on from this point and give you examples based upon faulty program data, the basic program data that decision-makers use; and I could take you through the OSHA program. I could take you through the Comprehensive Employment Training Act. I could take you through the Office of Federal Contract and Plans Programming. I could take you through the Wage Hour area and the Unemployment Insurance area and give you a litany here of issues that have been long-standing for decades. With the expenditure of the many many millions of dollars here, and the imposition of information costs on society which are not compensated directly, it would seem that we might be able to get at some of the answers to these questions, but in fact we cannot.

The EEO data we have, for instance, cannot measure the impact of the OFCC program, the Office of Federal Contract Compliance, either in gross or net terms. It simply can't do it.

We have an Occupational Safety and Health Program, and we do not know the nature of injury rates by occupation. It's just fundamental information that's lacking. There is a long-standing hypothesis that

minimum wages displace certain types of labor, and we are not able to establish at this point in time whether or not in fact that occurs.

What I'd like to say then, in summary, is that if you want to improve evaluation, and if you want to make evaluation operational, you must enforce the interaction of the program manager, the policy-makers, those people who gather data and those people who are presumably the information providers--the evaluators. If you don't do that (and obviously in practical terms you are going to do this at the staff level), if you don't insure this kind of interaction, I think you are simply wasting your time and wasting society's resources. Thank you.

MR. CAREY:

Thank you, Ernie. I am not sure that you have given us the elation we asked you for; but you certainly have given us some pretty solid things to think about.

XIII. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

MR. CAREY:

Now, you have seen this baseball team--nine players. They have done their bit. Let's take a few minutes to see whether you have any questions from the floor for Jim Stockdill or Ernie Stromsdorfer.

PARTICIPANT:

Charles Murray from the American Institutes for Research. This is relevant to the last two speakers, but it refers more to what I have been hearing all day about utilization, because one topic that has not come up is whether Government agencies are asking the right questions.[20] I see lots of RFP's with laundry lists of objectives, and I have met lots of program monitors who want to make sure that this topic and that topic and the other one is included in the evaluation. And I have almost never heard one tell me, "Don't worry about that because we can't do anything about it anyway."

From my perspective as part of a research company, it seems to me that the way to get an evaluator (who is not always that practically oriented anyway) to give you useful information that will get applied is not by hiring one who understands the political process in your bureaucracy. He shouldn't have to do that. He should be able to write, communicate clearly, have a good sense of what is practical and

---

[20] Editor's Note: Mr. Seeman did, in fact, raise this issue (see page 26 above); however, he appears to have been emphasizing the problem of asking the pertinent substantive questions about a program, as opposed to Mr. Murray's focus on practical questions (i.e., those questions for which answers provided by an evaluation can conceivably give rise to action).

what isn't. Above all, he needs from you a statement of the things which you can do which will take advantage of the findings he prepares. It's a statement to which I'd like your reaction, and the basic proposition is, you in the government aren't asking very good questions.

MR. CAREY:

Ernie, what do you think about that?

MR. STROMSDORFER:

I agree. If you don't have this interaction between the policy-makers, managers and the people who are supposed to provide information, you can't possibly ask the right questions. The policy-makers don't like to be put in a position where they have to formulate conceptual questions about their process and ultimate impact. The incentives are not structured in that direction with respect to the program managers. The big incentives are to invent a new program and get it funded. We have had different degrees of success in the Department of Labor in getting people to sit down and talk about these things. We have had very good success in the Employment Service, and we have very limited success in some other agencies. In one or two agencies where we talked to the program managers, they have simply allowed us to impose our value system on the program and on the questions that ought to be asked. And that's entirely wrong, unless, in fact, there is such a conceptual vacuum that it's better for us to impose our questions and our frame of reference rather than for no frame of reference to be imposed at all.

MR. CAREY:

I'd just like to comment. After I left Government I spent about five or six years as an officer of a fairly large consulting company. We saw the traffic of the RFP's. We had to. It was our business. But, as one who labored under the difficulty of having been in the

Budget Bureau for 25 years and one who thought he knew something about Government, there were times when I was appalled by the kinds of questions that Government agencies were asking outside consultants to address.

I remember one time when the Department of Transportation heard suddenly about a new Management-by-Objectives requirement from the White House. Over a weekend, they summoned in the blue-chip consulting houses, sat us all around and told us that what they needed in a relatively short time was for a consulting firm to define the objectives of the Department of Transportation. I was completely overcome--not with elation, but with concern as to how the hell the Government was being run.

MR. STROMSDORFER:

Well, Bill, that happens all the time. One of my predecessors did that for my shop, and the Urban Institite was brought in to tell us what we ought to think.

MR. CAREY:

I remember another occasion when the same Department discovered that the National Environmental Policy Act (NEPA) had been enacted with all of the various sections calling for impact statements, calling for revision of policy instructions and regulations and policy practices, operating procedures to conform to; a massive job, no question about it. But they turned to the consulting world for contract assistance in thinking out how the Environmental Policy Act applied to the tremendous array of different transportation programs in that Department. Again, we bid on that contract and the firm that I was with did indeed get the contract. I hope we were of some help.

But again, I was very much concerned that somehow that Government that I had so recently left just didn't have the internal capability to address those questions in a direct way, with only marginal assistance, perhaps, from outside houses.

I suppose that reflects my own sense of the proprieties and the way things ought to be done, and I guess I am not very objective. Is there another question for the panelists?

PARTICIPANT:

Sumner Clarren with the Urban Institute. I guess this question really has to do with how you organize to do evaluation. As I have listened, it seems to me (and I'll be making a caricaturization, too, I believe) that there is a difference between, for example, John Evans' approach--which is to have evaluations centralized, tightly controlled, featuring very prescriptive RFP's to purchase information to meet particular needs--and NIMH's view, as I see it, which says that evaluation, in a sense, is somebody else's business. Of course, NIMH wants to further knowledge, but they ask the mental health centers to get it; and the major requirement is that the centers send in an evaluation report every year. There are some general guidelines, of course, from Congress about the kinds of things the centers should measure, but the centers set their own priorities so that the design and a lot of the responsibility are both pushed down to the local level. These are two very different strategies for doing program evaluation; they are both <u>called</u> program evaluation at any rate.

I wonder whether it's an accident that these approaches have developed this way, or whether it represents something about the political origins of the programs. In other words, I guess I am

asking, is there wisdom in this kind of difference because it's taking into account something about the different contexts in which you both operate? Or is it just an accident?

MR. STOCKDILL:

I think we are both talking about the same thing. We believe strongly that if you are going to evaluate a program called Community Mental Health Centers and there are 600 of them out there, the only way your evaluation, using some sample of those centers, is going to be effective is if they have their own data collection system, are collecting valid data for their own evaluation purposes. So we began to feel very strongly after two or three years of a lot of these contract studies that we really had to improve the evaluation capacity out there in the field in order to improve the national capacity. I think we are both looking towards influencing national policy and national programs.

PARTICIPANT (CLARREN):

It seems to me that you have a very different strategy and maybe a different purpose.

MR. CAREY:

I am going to declare available and vulnerable not only Jim Stockdill and Ernst Stromsdorfer, but also Tom Kelly and John Evans. You can go at all four of them for the next few minutes if you so choose. Anything else?

PARTICIPANT:

Seymour Brandwein, Labor Department. You are reaching for a note of elation. I think we can be elated by some of the candor here, the willingness to recognize and acknowledge problems, although

I believe that many horror stories, even if accurate, ordinarily are a caricature that don't provide the overall picture.

My major concern is that we tend to mix up what evaluation can do, what it might do in some circumstances and at some times, and what it can't do inherently or in a particular decade. If we proceeded in that framework, I think we would be less likely to blame evaluation for not overcoming some fundamental problems of the sort that Ernie raised and that we really should not look to evaluation alone to resolve.

I was impressed with Stockdill's effort to pull out some of the sorts of things that can be done by evaluators. I think if we try to enlarge on those, and recognize that we are still in an infant activity, we might develop more of a basis for elation.

PARTICIPANT:

I am Paul Hammond, University of Pittsburgh. In the interest of proceeding in a constructive vein, I want to make a comment about what John Evans said and then make sure I do it in a way that will evoke some response from him. I want to suggest first that he offered us a nice complete process for evaluation that included gearing it in to a decision-making operation. It is impressive, and we ought to treat it seriously as one of the good examples to pay attention to.

Having said that, let me suggest that it works in part because he is dealing with a fairly stable constituency. I might even call it an organized constituency. I am not sure what it consists of.

MR. CAREY:

You might go even farther if you wished to.

MR. HAMMOND:

I am going to in a moment.

If one wants to look at the difference between evaluation operations that succeed and those that fail, one may find that the kind of political infrastructure of successful evaluation operations is going to be stable and may be organized in some sense. But the evaluation process then represents part of an interest process, and I am not sure that I like the good guys-bad guys version of Mr. Evans' story, perhaps because during some part of the time he is talking about, I was watching as an outside observer as some people under Richardson asked questions from the Secretary's Office that went to challenge the educational evaluation system, of which Mr. Evans is an important part, by saying, "Shouldn't we give the money to the students and get a market response rather than give it to the universities?"

Viewed from the Office of the Secretary, the effort to get an answer to that question wasn't very successful. Some of the reasons for failure may have had to do with people and stupidity--that is, competence and skill--but they also had to do with organizational processes, the fact that the information generation process (again, of which the Office of Education was an integral part), was mainly generating information that supported the status quo system--namely, channeling Federal funds through the universities, rather than through students.

I am suggesting that evaluation can work well if there is a consensus. I do not mean the kind of scientific consensus that Thomas Kuhn refers to. This is a different kind of consensus. It amounts to an organized, or at least an orderly, constituency. I am suggesting, that is to say, that an orderly constituency may be

necessary for supporting an institutional base for evaluation and research; and I'd be interested if John Evans agrees. Is he part of a process that depends upon a constituency-based consensus? And if so, well, is this as far as one can go with describing that process and accounting for the quality of its performance?

MR. EVANS:

Well, I hope that orderly constituency is not the hobgoblin of small evaluators' minds. Maybe I should just add a word, some historical background which others of you here may not be familiar with. I would certainly want to disclaim that the work that went on in Elliot Richardson's office had anything to do with any of the bad guys in my scenario. Quite the contrary, as a matter of fact. The particular effort Mr. Hammond is referring to is PEBSI, Program Evaluation by Summer Interns, which was an effort launched to do just what the acronym says. It did in fact fail, and one can analyze that failure from a number of points of view and a number of reasons.

MR. STOCKDILL:

John, that was an employment program; and looking at it from that standpoint, it succeeded!

MR. CAREY:

Continue with the objectivity, please.

MR. EVANS:

I come to the matter of evaluation in a fairly simple-minded way which says that basically, what we are talking about when we try to evaluate Federal programs is: are they effective? That is, do they achieve their objectives, objectives that are in the law, objectives that are <u>given</u>, despite some flexibility that must occur in the regulations which several observers have commented on.

We have a large $2 billion program in the Office of Education called Title I of the Elementary and Secondary Education Act which disburses $2 billion worth of money each year to states to form the grants where the purpose is to remediate the educational deficits of disadvantaged children. When I go down that track, I am very quickly led to the conclusion that summer interns do not have the competence and evaluation technology to answer that question. In order to answer that question persuasively so that one will want to form policy on the basis of it, and spend money on it, and make changes on it or not make changes on it, there is a need for a highly sophisticated kind of experimental design to determine whether disadvantaged kids who went into the program ended up later performing better than comparable kids who didn't go into the program. That kind of question is the basic question that applies to most social action programs and in my judgment must be answered with the evaluation technology that is appropriate; I think the PEBSI program was a clear example of the kind that is not appropriate.

So, to move from that point of your question to the other one about the established constituency, the only thing I would say there is this: I think basically the answer is yes, that the real clients of evaluation work that is carried out in Federal agencies are few. They are exeuctives, heads of executive agencies. They are the White House, the President and the OMB, and they are the Congress. Those are extraordinarily stable constitutents except insofar as individuals in the position change. Again, I think it's important that appropriate information should go to those people.

Of course, we also have the public; and that is different. I don't know whether I am sticking to your question or not, but I

would say that the kind of methodology or the kind of system that we have developed is one that, in our judgment, is best calculated, hopefully, to yield the least ambiguous, most defensible, and most relevant kinds of information with respect to program effectiveness that would be useful for decision-making to the several branches of the Federal Government.

MR. STROMSDORFER:

I'd like to comment on this statement of having an established constituency. I think it is this lack of a consensus or established constituency, for instance, which, in my judgment, has brought the Occupational Safety and Health Program to pretty much of a stalemate. This is curious because the law itself passed by an overwhelming majority (the law is an interesting law, too, because it does recognize, although not as clearly as I would like it to do, that there are costs involved in administering a program like OSHA and that there are likely to be some social conflicts arising out of your efforts to administer this law).

To repeat, the law was passed by an overwhelming majority, and the moment we undertook the effort to make the law operative, we came to a grinding, crunching stalemate. I don't understand quite what is going on here. My knowledge of the democratic system and of political science isn't great enough to encompass this. It's a curious situation.

Not only has the program come to a grinding halt, but our efforts to try to find out what is going on with what is happening are pretty well stymied, too.

MR. CAREY:

All things have to come to an end. The panel is at an end. The Chair retires and yields to our hosts, MITRE.

MR. GRANDY:

Thank you, Bill, and also my thanks to all of the members of the panel. I recognize that these presentations have taken somewhat longer than we anticipated. Judging from all of your perseverance here, however, I think they have been very helpful and illuminating. There are some common threads through them and also some very diverse ones.

We are going to take a coffee break, but we want to reconvene and continue our program until about six o'clock or as close thereto as we finish that part of our program. At that point, we will adjourn for our reception and dinner. Let's stop now for some coffee. If you would return promptly, it would help us.

THE RESEARCH PERSPECTIVES PANEL

I. INTRODUCTION: EVALUATING THE EVALUTORS

JAMES G. ABERT, Vice President for
Research and Development,
National Center for Resource Recovery

MR. GRANDY:

At this point in our program, we are starting the second phase which will continue until tomorrow morning. This is the discussion of current research experience from the perspectives of the researchers.

We are pleased to welcome Dr. James Abert who will given an introduction to this part of our program. Jim is currently Vice President for Research and Development at the National Center for Resource Recovery. He is a Mechanical Engineer with a doctorate in Economics. He was Deputy Assistant Secretary for Evaluation and Program Monitoring at HEW for two years, between 1969 and 1971.

In between the time he left HEW and today, he has studied and published widely on a variety of topics. In addition to his Government service and his industrial work, he has been involved in a number of policy study committees for the National Science Foundation, for the National Academy of Engineering and for the National Academy of Sciences.

It's a pleasure to welcome him. He is going to talk on evaluating evaluators; hopefully, I think, in terms of Jim Stockdill's warning, only one level deep.

MR. ABERT:

Thank you very much. Some of you may know that the National Center for Resource Recovery is concerned with refuse recycling. I

have been told on several occasions (more often, the further I get from Washington, D. C.), that a couple of years of HEW is really good preparation for a career in garbage.

Both in Government and in the research community, it should be clear by now that the term evaluation lacks precise definition. Among producers and users, there is wide variation as to what constitutes evaluation and how it differs, if it does, from (among others) field experimentation, demonstration-research projects and, to choose another term, action-research programs.

Exact definition, however, is probably of little importance. Regardless of the exact meaning of the term, it appears that evaluation has become somewhat of a fad, if not yet an entirely proven, integral part of the management process.

I think it is important to ask at this time if the resources devoted to evaluation are a valued activity in the constant search to improve the efficiency with which public sector funds are spent. This is not to suggest that a definitive answer to the question can be given.

I would say that throughout the Government, the foundations for evaluation laid some years ago have grown into a full-fledged evaluation emphasis. I have chosen the word emphasis carefully, and it is to stress that the development and institutionalization of an evaluation program is an evolutionary process. It is not done overnight. Indeed, it is not done in a year or two. How long depends on the interest and determination of those responsible for its direction and the support given to its growth. To graft it to a hostile bureaucracy requires both toughness and tender loving care. It does not "take" easily.

The stakes are high, not only because of the employment generated within the Government but its effect on that segment of the private sector which responds to the RFP's to do the evaluations. I have heard evaluation characterized as Federal aid to contractors, and to some extent that is true.

However, the real stakes are where the big bucks are. Evaluation can become, perhaps inevitably is, a political device which can be used to promote support for an advocate's program or reduce enthusiasm for an opponent's proposals.

Evaluation is important in other areas as well. It provides the financial incentive for academicians to train their intellectual ordnance on the target of improving the management of public funds. Some may argue that they often fail to find the target. Perhaps they fire with biased sights, or perhaps the target itself is poorly understood by those in the user community whose articulation of what mark was to be hit is often only clear to them after the fact of the evaluation.

Finally, evaluation provides the wherewithal to expand the general knowledge base in areas where the more traditional data collection services have not ventured. At the least, the social sciences should have seen and should continue to see more dissertations in what might be called the "grand design."

Putting this aside as a spillover benefit, and presumably it is a benefit, the basic question concerning the valuation of evaluation is "Do evaluation outlays produce greater efficiency in program output than the costs of the evaluation efforts?"

In general, short of saturation, more information is better than less. Yet there are costs involved. Are the likely improvements in program targeting and management worth the cost of data collection and analysis? Evaluation can cost more than it is likely to save, although the definition of "save" is a problem here.

There is a more subjective side as well, indeed even emotional--so emotional that evaluation can become counterproductive. This is particularly true when one begins to evaluate in earnest, where only lip service has been paid to this function in the past.

As you know, the setting of program objectives and the choosing of evaluation criteria are in themselves a very emotional undertaking. Program managers generally are not anxious to do it. In fact, trust, confidence, honor and many of the more noble aspects of life seem to be strongly challenged by evaluation.

The tools for estimating the worth of policy-related information are primitive at best. Much of the information obtained simply helps the program manager to understand his program better. To relate this information in some casual way to program improvement and then to further measure the value of this improvement appears to be beyond today's practice.

Partially for these reasons, the usual chronicles of evaluations accomplishments--a successful study or two offered as evidence of the achievements of the evaluation program--often leave the listener with a feeling of "Well, maybe the expenditures on evaluation have been worthwhile; and again, maybe they haven't."

Has progress been made? At the outset, it seemed that many felt an evaluation program should appear fullblown. Of course, this has not happened. The formulation and implementation of a viable program is a step-by-step process. One builds on what one has accomplished in the prior period. One does not grasp for options that have low probability of being achieved. The need to set reasonable sights and to plan for evolutionary growth with many mid-course changes does not seem to have been appreciated fully either at the outset, or now. Also, it is usually not present when observers of evaluation programs, no matter how objective they may claim they are, attempt to evaluate evaluation.

There is still much to be done. The key to the future growth and acceptance of evaluation is the development of recognized approaches to the conduct of evaluation, in particular to establishing the reliability of the judgments made by field staff.

Of course, it is necessary initially to obtain and to maintain high-level support. Because evaluation's image is that of uncovering or demonstrating the negative, it is generally only grudgingly and reluctantly accepted by those on the receiving end. While the degree of support of the evaluation activity can be reflected in a variety of ways, the position of the evaluation office in the organizational structure will be a principal indicator. Clearly, if such units are directly linked to principal decision- and policy-makers, the possibilities of influence will be noted throughout the organization.

Along the same line, evaluation, in my view, should be legislatively mandated and treated as a program in its own right including a mandated budget.

In addition, thought must be given to evaluation in the structuring of operating programs such that more of them can, in fact, be evaluated.

Finally, program evaluation must hew to a nominative approach that forces judgments as to the accomplishments, or lack of accomplishments if such is the case, of the program being evaluated. In general, the research paper suitable for publication and useful for promotion does not fit the bill.

Time will bring with it a greater appreciation for the real-world context of evaluation. This must be so, or the evaluation parallel will be that of the formal discipline-focused research program, lodged far down in the agency, far from the policy arena.

Looking to another facet of the evaluation picture, it is too soon to tell if the political process has been sufficiently sensitized to allow evaluation to continue with its evolutionary growth. There are still many hurdles to be overcome, not the least of which is institutionalizing evaluation requirements, procedures and dissemination to the extent that past lessons are not relearned by each succeeding change in department and agency management.

Only time will tell whether evaluation lives up to the reasonable expectations of its advocates or turns out to be a relatively short-lived but expensive experiment. Thank you very much.

MR. GRANDY:

Thank you, Jim. In your remarks I think there are some reinforcement of comments of some other speakers this morning. Also

something of a challenge to this audience to help facilitate the institutionalization process so that evaluation becomes solidly enmeshed in the fabric of public program management.

II. STUDENT ATTRITION AT THE FIVE FEDERAL SERVICE ACADEMIES: AN IN-DEPTH AUDIT

CHARLES W. THOMPSON, Assistant Director,
Federal Personnel and Compensation Division, and

JOHN K. HARPER, Acting Director,
Systems Analysis Group
Federal Personnel and Compensation Division
General Accounting Office

MR. GRANDY:

Next on our program will be the presentation of our first research paper. This will be done by two gentlemen from the General Accounting Office, Charles Thompson who has long been a staff member at GAO, and his colleague, John Harper. Both are in the Federal Personnel and Compensation Division. Their report, as you see from our agenda, deals with Student Attrition at the Service Academies. The document is available, displayed with other literature out in our anteroom.

I think Mr. Thompson is going to speak first, and will then turn the discussion over to Mr. Harper.

MR. THOMPSON:

I'd like to spend a few minutes discussing the problem of attrition that we faced, our general approach to addressing the problem, a few of the more significant findings, our recommendations and some of my perceptions as to the factors which may have influenced their utilization.

The military academies exist primarily for one purpose--to develop career military officers.

Even though the academies account for only about 10 percent of the initial grade officers acquired by the military services, academy graduates are nonetheless considered among the more highly desirable officers.

To the extent that large numbers of students who would make good career military officers leave the academies before graduation, the effectiveness of the academies' program becomes questionable.

In recent years, attrition at the academies has been high, and it has been increasing, and these increasing trends, particularly at the Air Force Academy, prompted Senators Birch Bayh and William Proxmire, as well as other members of Congress, to request a GAO study of the problem.

Figure 2, below, will give you a better sense of what their concern was.

For four of the five academy classes which graduated in either 1974 or 1975, attrition reached near-term record levels. For example:

- The Air Force Academy graduating class of 1975 had a 46 percent attrition rate, the highest in its history;
- The Military Academy reached an 11-year high of 40 percent attrition;
- The Naval Academy, a 12-year high of 39 percent attrition;
- The Merchant Marine Academy, an 11-year high of 48 percent attrition; and
- The Coast Guard Academy had 46 percent attrition.

In light of these statistics, there were serious questions being raised as to whether the academies were adequately accomplishing their mission. When we add the additional consideration of costs--over $100,000 per graduate--the issue becomes not only one of program

FEDERAL SERVICE ACADEMIES
ATTRITION RATE BY GRADUATING CLASS YEAR

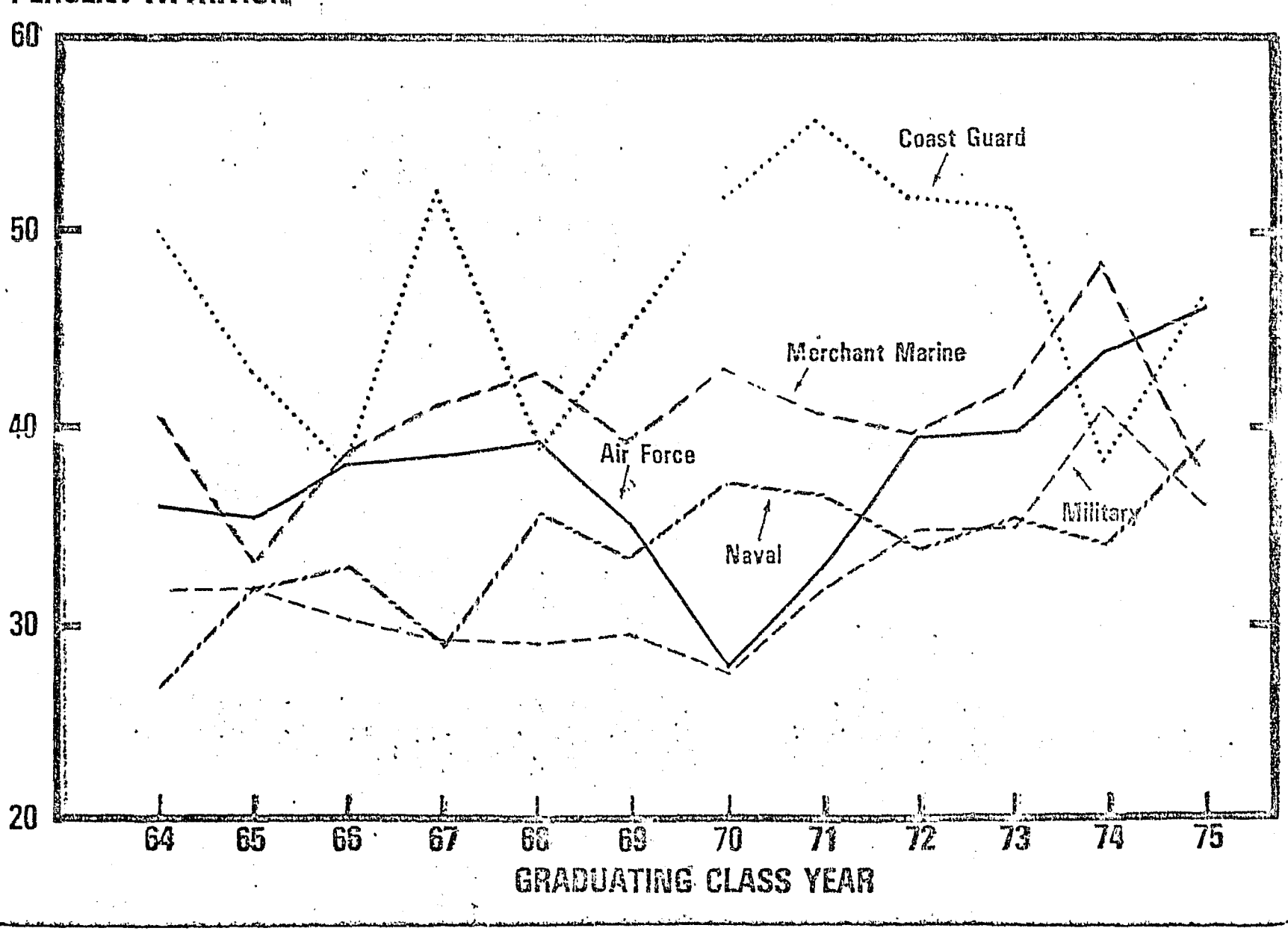


FIGURE 2

effectiveness but also one of program affordability. If attrition could be reduced, the academies could provide more graduates at a more affordable cost. Without reduced attrition, alternative sources of officer acquisition become more plausible and more attractive.

With this in mind, our study objectives became two-fold--first identifying those factors which contribute to the high attrition rates, and, second, proposing program alternatives which we believed would permit the academies to reduce their attrition rates without degrading the quality of the graduates.

We recognized at the outset that some attrition is inevitable and desirable since selection of only those who would make good career officers is unrealistic. Attrition, therefore, serves as a desirable screening device for those students who do not measure up to the standards considered essential to the military profession. Yet, our data suggests that, in addition to weeding out those whom the academies felt were undesirable, they were also losing many potentially good career officers. In fact, one academy superintendent estimated that 20 percent of voluntary dropouts were potentially good career officers.

We felt, therefore, that if we could identify those major factors contributing to the student attrition and recommend changes to them without decreasing the quality of the output, we would be making a contribution to improving the effectiveness of the academies' program at a more affordable cost.

Let me very briefly review for you our approach to the attrition issue, for it is the acceptance of this approach and the steps that were taken to increase acceptance which determined, at least in part, the acceptance of our results and the extent of implementation of our recommendations.

Figure 3 below, shows a rather simplified version of the model we adopted to identify the factors contributing to attrition.

Conceptually, we viewed attrition as resulting from the interaction of three distinct influences: (1) the characteristics that students bring with them to the academy, such as abilities, commitment and expectations, (2) the effect of the academy environment on the students, such as the quality of the academic and military programs, and (3) the external environment which affects students while they are at the academy, such as national economic conditions in general.

Through a rather extensive review of the existing research on attrition, as well as through discussions with current and former academy officials and students, we identified those factors within each of these three areas that could potentially contribute to attrition. These factors, then, formed the basis for our data collection efforts.

Our primary data collection source was a questionnaire we developed and administered to over 20,000 current and former academy students. In addition, we obtained extensive data from academy records and from an annual survey of incoming academy students administered by the American Council on Education. In total we collected or obtained over 500 specific items of information on each student which we hypothesized were related to attrition.

Because of apparent differences in the academies' environments and in the students who go there, we decided to perform separate analyses of each academy.

Further, within each academy, separate analyses were made for each of the three timeframes--the first summer preceding the fourth

ATTRITION MODEL

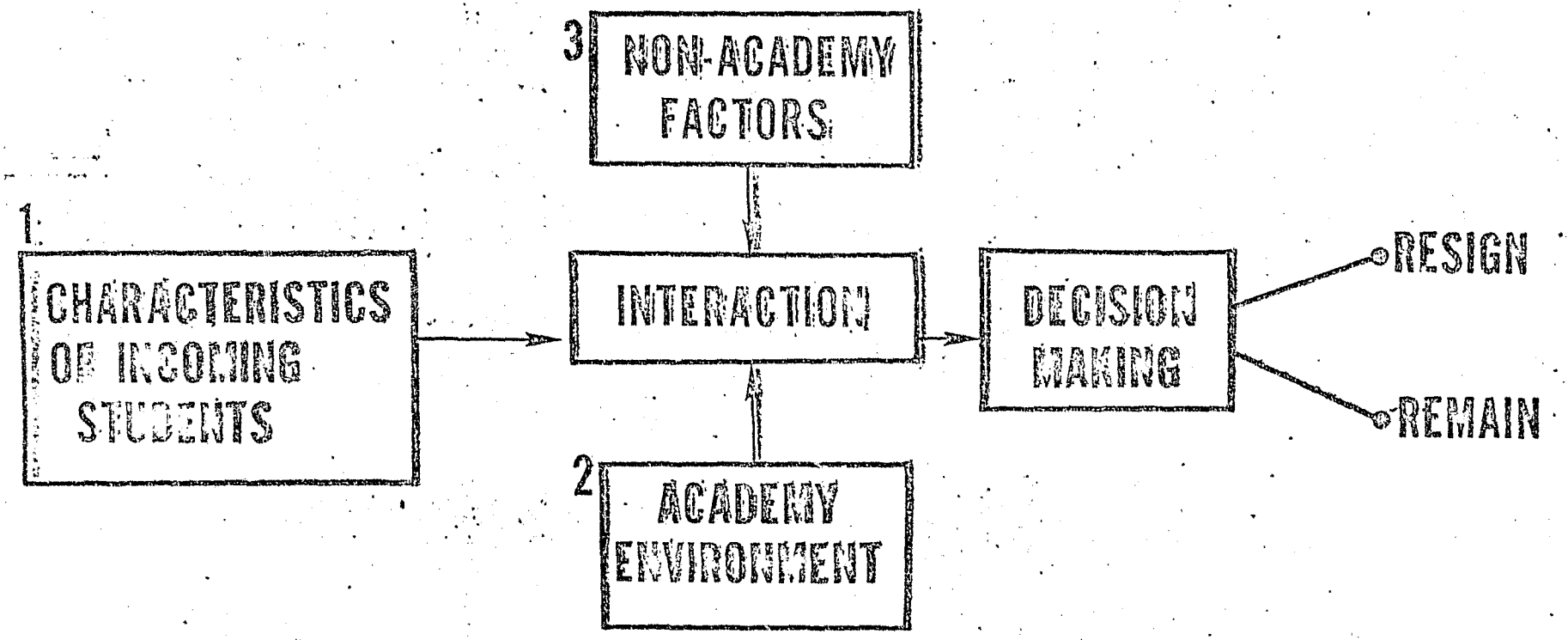


FIGURE 3

class academic year (this is normally the first two months that a student is at the academy); the fourth class or freshman year; and the third class or sophomore year. These three timeframes were chosen because about 85 percent of all attrition occurs during these first two years and because our prior research indicated there to be different reasons for attrition depending on the timeframe.

Let me now briefly discuss some of our findings to give you a perspective of what they were like and the recommendations we made from them.

In general our recommendations tended to fall into three categories: First, major changes to academy policies, practices or tradition. These tend to be rather high-risk changes, in that if they were implemented and later proved to be wrong, they could have a major detrimental impact on the academies' mission. Second, relatively minor changes to policies or practices. They tend to be low-risk changes. And third, recommendations for further research or redirection of research.

Let me illustrate by discussing a few specific findings: we found that one of the most important factors related to attrition during the students' first few months at the academy is their initial level of commitment at the time they entered. Those students who have lower levels of commitment have significantly greater probability of dropping out.

Our measure of student commitment was made up of a number of questions which the students answered when they entered the academy. These concerned the chances they would transfer to another college before graduating, drop out of college temporarily or permanently,

change their career choice, or get married while in college. Each of these actions almost always requires the student to leave the academy.

Those who dropped out saw their chances of doing each of these things to be significantly greater than those who stayed. Figure 4 below illustrates this point. It shows the responses of first summer dropouts and current students about the chances they would transfer to another college before graduation.

At the Air Force, Military, and Coast Guard Academies, 35, 31 and 46 percent respectively of first summer dropouts stated at the time they entered that there was a "Very Good Chance" they would transfer to another college. Whereas only 2, 4 and 6 percent respectively of current students made this response.

This initial level of commitment is extremely important. There have been leaders at the academies who adopted a philosophy that if a student doesn't want to be at the academy, then the academy doesn't want him. And their programs, especially during the first summer, were designed to test, and I want to emphasize the word test, a student's commitment. However, our study suggests that this philosophy may have driven some good students out.

It's my view that the academies failed to adequately recognize that low commitment is typical of individuals at this age. For example, the next figure gives an indication of this low commitment as it relates to the academies (see Figure 5 below).

For the total class which entered in, for example, 1973--this would be the far right bar on each chart--between 43 and 58 percent of students stated that there was some, or a very good, chance they

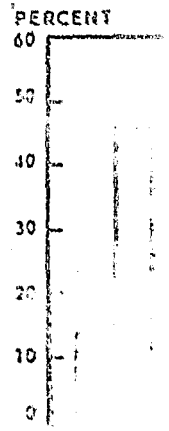


FIGURE 4

PERCENT OF ENTERING STUDENTS WHO BELIEVED THERE WAS "SOME" OR "A VERY GOOD" CHANCE THEY WOULD CHANGE THEIR CAREER CHOICE

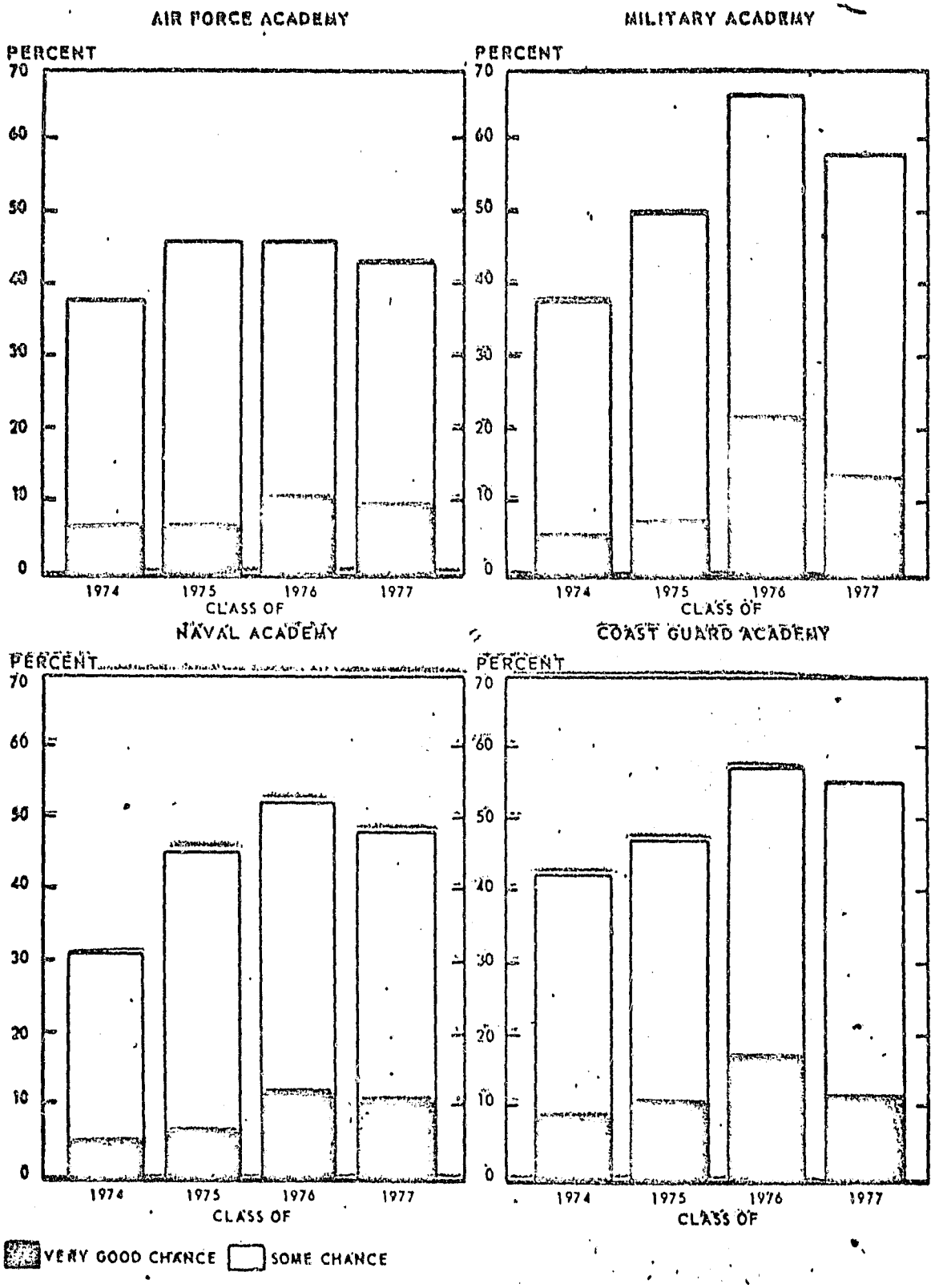


FIGURE 5

would change their career choice. The point I am trying to make is that it is better for the academies to make the assumption that students are not highly committed and design their programs to develop commitment, rather than merely test it.

Generally, the academies agreed with our finding on the importance of commitment to retention; and at least the Air Force Academy instituted an intensive reexamination of their first summer program in an effort to make it more commitment-developing rather than commitment-testing.

However, the responses from all academies were very mixed on the extent to which they agreed to institute specific changes to practices which appeared to be more commitment-testing than commitment-developing. For example, we found that the requirement to memorize and recite trivia, such as sports scores and titles of movies, the heavy emphasis on drills and ceremonies, and the heavy emphasis on creating stress were directly related to attrition.

The more minor of these changes were readily accepted. For example, the Military Academy reduced the level of drills and ceremonies by 35 percent, with further reductions planned. The need to reduce the memorization and recitation of trivia was also generally accepted.

On the other hand, a more major change, that is, the need to review and possibly modify the extent of stress in the environment, was not accepted. In fact it was strongly defended as necessary.

I'm not suggesting that the academies should have accepted all of our findings and made immediate changes. The point I'm trying to make is that the degree to which a finding is accepted and acted

upon is, to some degree, a function of the potential risk-level of the change. And the greater the risk, the more the decision-maker will require additional supportive data before a change is made, particularly if the change is contrary to his predisposition.

Therefore, while the results of some of our findings were not immediately acted upon, they did provide an additional source of information, which, when combined with other supportive studies to follow, will hopefully result in a critical mass and cause a change.

We should not always be disappointed when high-risk type recommendations are not acted upon. We don't necessarily have to live with the consequences.

In closing, let me summarize what I perceive to be some of the factors which influenced the use of our evaluation results.

First, use is, at least in part, a function of the extent to which the decision-maker has confidence that the results are valid; and this again to some degree is a function of the soundness of the approach and the clear, understandable link between the approach, the results and the conclusions and recommendations. We can increase acceptability and use by involving the decision-maker, or subordinates whose opinion he respects, in the process from beginning to end. Recommendations for change, particularly major change, should not come as a surprise at the end.

Second, if we can involve other outsiders of the group doing the study in the study process--again ones whom the decision-maker respects--we provide an important secondary group to whom the decision-maker can look for confirmation of the conclusions and recommendations.

Finally, don't expect that all recommendations will be acted upon. The higher the risk, the less chance that change will take place from the results of one study, no matter how sound. Also, the greater the decision risk, the greater the need to bring the decision-maker and others along with the study.

It is my personal view that researchers or program evaluators can and need to have more interaction and communication with the ultimate decision-maker. If we are to maximize the chances of results implementation, we need to build a greater sense of trust between the decision-maker and the evaluator--trust in his methodology, trust in the validity of his conclusions and the soundness of his recommendations, and perhaps, most important, trust in the evaluator himself. Thank you.

I'd like to turn the discussion over to John Harper, who will further discuss some of our findings and some recommendations.

MR. HARPER:

I will talk about two factors which seem to have affected the extent to which findings from our study could have been and, indeed, actually were used as a basis for policymaking. Let me stress that this is a personal view. Others might well have seen different factors as crucial in determining the extent of utilization.

The first factor was the context in which the study was done. I would like to talk about that context in terms of power relations among the principal actors in the study (see Figure 6 below). I want to do that because it's my feeling that those relations: (1) made the study possible, (2) partially determined how the study was done, and (3) affected the extent to which it was utilized.

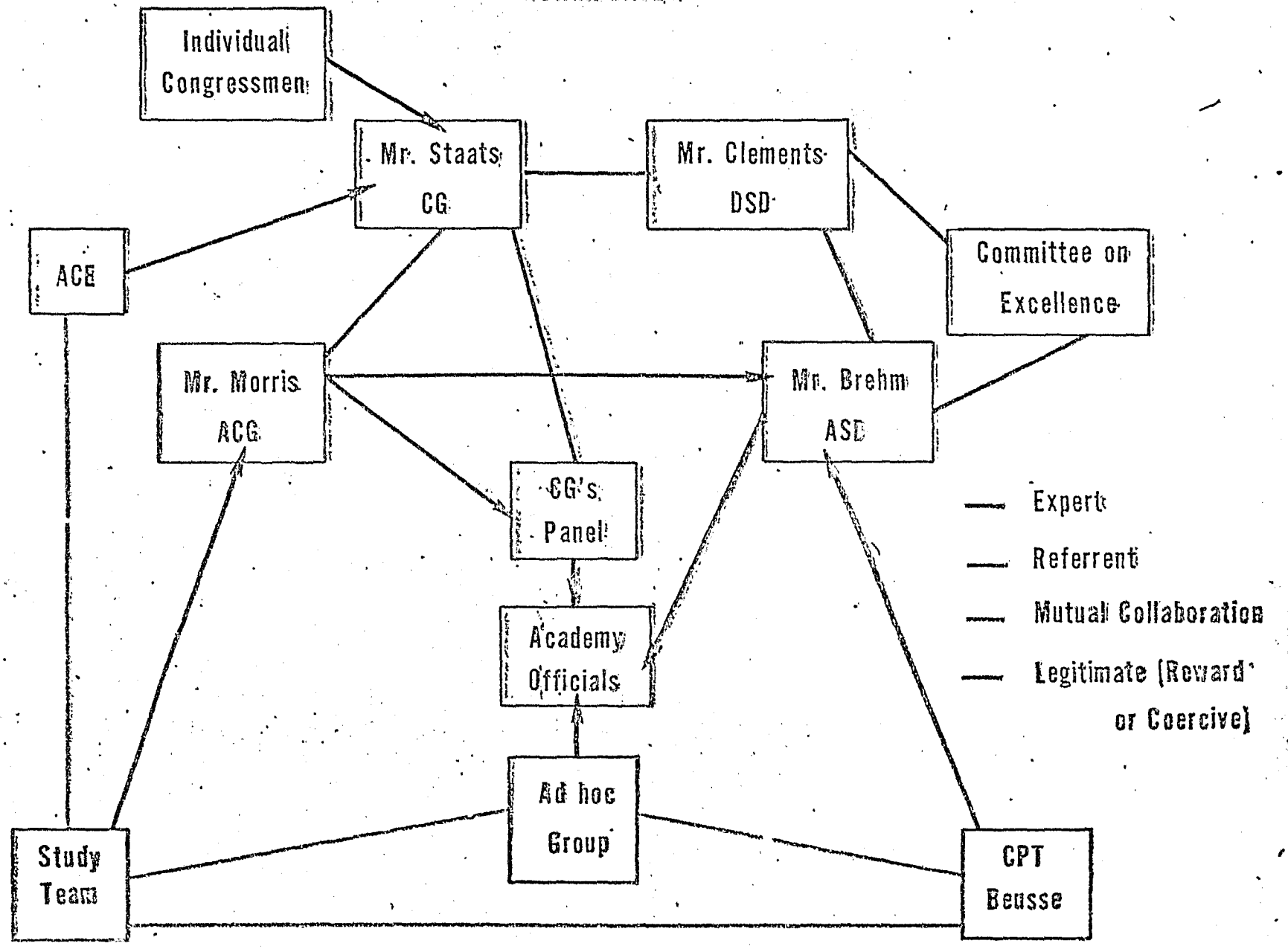

POWER BASES
Individual Congressmen
Mr. Staats CG
Mr. Clements DSD
ACE
Committee on Excellence
Mr. Morris ACG
Mr. Brehm ASD
CG's Panel
Academy Officials
Ad hoc Group
Study Team
CPT Beusse
Expert
Referrent
Mutual Collaboration
Legitimate (Reward or Coercive)

FIGURE 6

There are at least four types of influence, or power, one person can have over another. The first type is expert power which exists when the division of labor in an organization produces groups or individuals with specialized knowledge or expertise necessary to accomplish the organization's primary mission.

The second type is referrent power and is based on the extent to which one person identifies with, or is attracted to, another person because the other person behaves or believes like the influenced person.

The third type of power flows from formal organizational relationships which permits someone to dispense sanctions and rewards based on shared norms.

The last type of power is mutual or collaborative power where the direction of influence alternates between actors.

With these brief definitions in mind let me give a personal view of the principal actors involved in our study and the types of influence they exerted over one another. These remarks are limited to the military academies (Army, Navy and Air Force) because they were the opinion leaders for the other academies in this study.

As Chuck mentioned earlier, the study initially requested by several members of Congress was of factors related to attrition at the Air Force Academy. Attrition had risen dramatically there, and the Superintendent had made a number of hard statements about the institution's lack of concern over it. Several of those who requested the study were perceived by some in the Department of Defense as holding unfavorable attitudes toward the military.

For various reasons, Mr. Staats, the Comptroller General who has headed GAO since 1966 and who was Deputy Director of BOB for many years before that, decided that the study should not have a limited focus but should extend to all of the service academies.

About the time this decision was made, Mr. Staats was in touch with the then Secretary of Defense, Mr. Schlesinger, and Mr. Clements, the Deputy Secretary of Defense. Both of these individuals had been concerned with civilian oversight of DOD's education programs. This concern led to creation of the Committee on Excellence in Education, composed of Mr. Clements and Mr. Brehm, the Assistant Secretary for Manpower, as well as the Service Secretaries. The academies were a principal item on the Committee's agenda.

Mr. Staats was also in touch with a number of members of Congress who expressed reservations about the benefits to be gained from this study. Senior officials at several academies also expressed reservations about the study.

Recognizing the sensitivity of the issue being addressed, Mr. Staats and Mr. Morris, the Assistant Comptroller General, decided to bring together an outside panel to consult with GAO on the study. Mr. Clements suggested a number of former, high-ranking academy officials as candidates for the panel. To add balance, Mr. Staats solicited names of civilian academic administrators from the President of the American Council on Education. The panel which was established consisted of Chancellors of the Universities of Texas, Illinois, and Pittsburgh, and the President of Tuskegee Institute; Vice Presidents of Harvard, MIT, Stanford, Michigan, and Tulane; and former

Superintendents of each of the academies. The other members of the panel--there were 17 in total--were no less important or illustrious.

A number of the civilian administrators had held high-level positions at the academies. For instance, the Chancellor of the University of Pittsburgh had been Chairman of the Social Science Department at West Point and Chairman of the Economics Department at the Air Force Academy.

We met formally with this panel on five occasions over a two-year period and met informally with individual members several times during that period.

Prior to the first panel meeting, the project team presented a proposal for this study to Mr. Morris, who had served as an Assistant Secretary of Defense on two separate occasions. Mr. Morris liked what he heard and communicated that feeling to Mr. Brehm.

The proposal was also enthusiastically received by key panel members at that first meeting. The panel gave the study a certain kind of legitimacy. It also forced us, as researchers, to keep our feet on the ground and it served as a vital communication link to senior academy officials. Meetings of the panel were held at each of the military academies and their senior officials participated in the meetings. This opportunity for them to express reservations about the study to such an illustrious group and to have those reservations moderated--when combined with the informal conversations which occurred between senior academy officials and some of the panel members, I believe, helped to overcome the resistance mentioned earlier.

Mr. Staats' and Mr. Morris' interactions with senior DOD officials added another kind of legitimacy which became important in overcoming particularly strong resistance by academy officials at certain other points of the study.

At the more mundane level, the study team was strongly influenced by the work of the Office of Research at the American Council on Education. The design of the study imitated the "input-output" model which had characterized ACE's research on college impact since 1968.

The research team was fortunate to have as a liaison in DOD a Captain with academic experience in organizational behavior and work experience in survey research. His expert influence was helpful at upper levels in DOD, and his collaboration was helpful with an ad hoc group we had formed to provide us with technical assistance. This group was variously composed of mathematicians, psychologists, and management scientists from the academies; computer scientists and researchers from the military personnel labs; and manpower program managers from the service headquarters.

We met formally a number of times with this group and informally with some of its members. The circumstances surrounding our second meeting give some example of the types of influence at work on this level of the project.

Prior to that meeting, the GAO study team had developed a pool of questionnaire items, and had discussed the study design and hypotheses behind each item with its own field teams. Those field teams had returned to the academies with the questions typed--and not very neatly--one to a page, to discuss them with senior academy officials and ad hoc group members. The teams had been instructed to emphasize

that the questions made up a first draft item pool, and that we were primarily concerned with whether hypothesized causes of attrition had been adequately sampled.

The scales for some of the questions were not balanced, and a number of items were clearly biased against the academies. During the time the field teams were discussing the item pool, we corrected many of these deficiencies. But we made a mistake. We typed the corrected questions one after another in survey questionnaire format. We added response boxes which had not been there before for each question and, in short, developed a fairly professional-looking questionnaire--even for a draft.

The ad hoc group began arriving at the Pentagon from all over the country at 8:00 A.M. on a Monday morning for what was to be a one-day strategy session among themselves before meeting with the GAO study team. Many of them had been given strong marching orders when they left the academies. Well, when they were given the new draft of the questionnaire with its completed-looking appearance, it met with strong resistance. By 9:30 that morning, the meeting with the GAO team had been cancelled; and talk was that the ad hoc group had been dissolved, the academies would not let their students participate in the study, and any study we might be able to do would not be considered legitimate by the academies.

Needless to say, there was a great deal of sideways and upward communication. Chuck and I communicated with Mr. Morris. Captain Buesse communicated with Mr. Brehm. Mr. Morris and Mr. Brehm communicated. And Captain Buesse and I spent several days discussing hypotheses and response scales.

I don't know whether it was reason, or power, or something else, which prevailed; but the ad hoc group did meet about two weeks later. In a hectic five-day meeting, the study team established its professional credibility and convinced the program managers that we weren't trying to give the academies a bad name.

As time went on, the ad hoc group supported the study more and more to academy officials, although they were never blind to its technical problems.

I have spent a good deal of time talking about dynamics at work during the study because they changed by the time it was done. They changed because the actors changed. By the time the reports were issued, Mr. Morris and Captain Buesse and many senior academy officials had left. I can't say how these changes affected utilization, but I do feel they were important in developing something which could be utilized. I also believe that the various advisory groups served as a vital, independent communication link between us and the academies; and further, where our methodology and findings were accepted by them and communicated to the academies, the probability of implementing those findings was increased.

The second major factor affecting the utilization was the intractable nature of some of the technical problems in inferring causality, interpreting factor scores, assuming a certain model, and nonresponses.

We had no control over treatments, and random assignment was out of the question. For that matter, we did not know enough about critical variables to design an experiment. Moreover, we did not have the time to conduct a panel analysis which would help us infer the direction of dynamic relations. Finally, the limited number of

academies precluded drawing meaningful conclusions about objective organizational characteristics such as ACE had done in assessing college environment impacts.

We were left, for the most part, with a post-hoc, correlational study based on self-reports of academy experiences and evaluation. In short, a weak foundation upon which to base recommendations for change.

We collected a great deal of information on each student and dropout. At the prodding of the ad hoc group, we performed a series of factor analyses on the data.

For those of you who have done factor analyses before, let me say we learned something. The computer-generated factor scores were occasionally uninterpretable when one compared item validities with factor validities. Some of the factors were accounting for negative variance. And finally, the structure of some of the factors made it difficult to develop recommendations.

We assumed the general linear model throughout; and, perhaps as a result, the size of our correlations was not overly impressive.

Finally, while the rate of questionnaire return by dropouts was high (73 percent), it was not perfect. ACE conducted analyses of the non-respondent characteristics and could not conclude that they differed from the characteristics of those who did respond. By the same token, we could not conclude that the two groups were necessarily the same in terms of academy experience and evaluations.

We recognized all along that these limitations existed, and we candidly stated in our final report "that a correlational study (as

ours was) does not establish clear cause-and-effect relationships and that surveying student perceptions after the fact presents special problems of data interpretation. Alternative interpretations exist."

We tested the validity of those other interpretations as best we could, but admitted to not being able to recognize or test them all. Therefore, we went on to say "Because alternative interpretations are always possible from survey data of the type we collected, our conclusions and recommendations have been stated cautiously."

Despite these limitations, we felt we learned some things from our study. Probably the principal reason is that there was a research base on which we could build.

Several of the academies had been doing attrition-related research for years. We collected all of the studies that could be identified and focused on 84 of them for detailed analysis and synthesis.

These studies left us with two impressions. First, very few of them had to do with the environment at the academies--far and away, the majority had to do with the relationship between characteristics at entry and attrition. And second, perhaps only one of the environment studies could be considered to possess what Stanley and Campbell refer to as "internal validity"--the sine qua non of scientific research.

We found the studies useful, nonetheless, because they explored dimensions of student characteristics that we did not explore. Some of what first appeared to be anomalous responses in our questionnaire--i.e., dropouts responding the way we hypothesized current students would, and vice versa--became interpretable only when we considered the implications of those entry dimensions the academies had explored.

Let me give you an example. At West Point, we found that those who stayed were less certain than those who left about their responsibilities and about what officers or upperclassmen thought of their performance. Similarly, those who stayed reported being bothered by having too little authority and responsibility. Also, their view of leadership was to have upperclassmen encouraging them to give their best effort and maintain high standards of performance. Dropouts, on the other hand, had a view of leadership as support from classmates.

These findings became interpretable (because we were asking questions about the environment) only when previous academy research on personality characteristics was viewed in light of the intensely competitive environment of the academies. That research indicates dropouts are largely non-competitive and are not achievement-oriented. They appear to have higher needs for affiliation and affection. Those who stay are concerned about achieving in terms of a standard of excellence, and are more independent in their interpersonal relations. Clearly, role ambiguity and not feeling enough responsibility would be bothersome to such people.

After arriving at this interpretation, we suggested that West Point might want to reexamine the amount of stress and competition in its environment. The Academy and DOD didn't like that suggestion. They pointed out that the stress and competition simulated what graduates would face on the battlefield where they would be responsible for the lives of others. We allowed as how this argument had appeal, but questioned what it meant with respect to other officer acquisition programs where students do not experience the same level of stress and competition 24 hours a day, 7 days a week, for 4 years. I don't think that any effect the reasoning out of the implications of the argument

might have had can be separated from the effects which the recent cheating scandal and Congressional interest in academy competition might have had.

As Chuck mentioned, where there was a predisposition by senior academy officials to accept our findings, they were in fact acted upon. The amount of drills and ceremonies and rote-memory of trivia were reduced, and some extracurricular activities were instituted.

However, where we challenged deeply-ingrained attitudes about the academies, there was strong resistance to our findings. The competitive environment was one such area. Another was the finding that dropouts did not perceive the educational program as having the high quality which the current students did.

The possibility of longitudinal research was precluded by the steps we took to insure confidentiality. Such steps drive up the cost of this type of research because you can't amortize the cost of design and data collection through repeated measurements. Nonetheless, we believe our study does add to the academies' fund of knowledge. But more importantly, we are an agency of the Congress; and ultimately our work should feed into their decision-making. In this case, it did. The Senate Committee on Appropriations used information from our study as one justification for recommending closure of the academy prep schools. The Committee also expressed an intent to critically review the academies' actions regarding our recommendations, and specifically with respect to the competition in the environment--a finding with which they agreed. Thank you.

MR. GRANDY:

Thank you very much, gentlemen.

III. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

MR. GRANDY:

Let's take a few minutes at this point and see if you have some questions of these speakers about their work or their presentation.

PARTICIPANT:

I'm Joel Garner with LEAA. I have a few questions, one of which directly relates to your study. First, who do you consider was your primary audience and who in effect set the objectives for the study? Clearly you work for the Congress and two specific Senators asked for this study. But it seems from your discussion that you worked primarily and directly with the service academies and with the Department of Defense, and that they, in the operation of the study, became the primary users or individuals who set the objectives of the study or of the evaluation. The second question is one that other people have raised but never answered, and the question is, should Congress in either legislation or requests to GAO be required to set specific evaluable objectives? If we can't expect agency administrators to do this, maybe we should expect Congress to do that and only do evaluations when Congress is very explicit about its objectives.

MR. THOMPSON:

I don't think there is any question our primary focus was to the Congress. We work for them, and we report to them. But we also recognized that we have other potential points at which we can have the results of our work implemented. If we can work with the academies throughout the job, still maintaining our independence, and get them to appreciate what we are trying to do, the objectiveness of what we are trying to do, they are much more likely to act on their own

to implement some of our recommendations. I think at least to some degree, that was the case. It was clearly the requests of Senators Proxmire and Bayh which set the scope for our work.

MR. HARPER:

What we tend to find is that the process of doing our work oftentimes achieves whatever objectives we might have hoped to achieve by producing a report because, at the point in time when the agency gets the opportunity to comment on the drafts of our report, my experience is that the agency likes to say, "we have already instituted action to correct whatever it is you have found."

MR. GARNER:

Did Senator Proxmire and Senator Bayh use your report in any way? Did they do anything with it?

MR. HARPER:

Let me just mention that we wound up ultimately on this job with something in the neighborhood of 12 to 16 requests from Congress, letters from Congressmen asking us to do work in this area. One of the spinoffs of this job was to look at the training programs in terms of the amount of harrassment that was going on. We issued a separate report on that.

MR. THOMPSON:

Frequently on a request from someone like Senator Proxmire, the request comes not from him as an individual, but as a member or Chairman of a subcommittee. At the completion of our review, the Comptroller General testified before the House Subcommittee on Legislation and National Security, Committee on Government Operations. It's really that Subcommittee that has the power to act.

One other point, I think, that is important is that very frequently we find that change takes place before the study is finished. Just the pressure of the Congressional inquiry, just the pressure of GAO looking at the problem is enough to get the academies or to get the agency to begin seriously thinking about it and begin to make changes. What we found was that shortly after we got in, attrition began to come down. Whether it was as a result of our work or not, I don't know. It's nice to think so.

PARTICIPANT:

Paul Hammond, University of Pittsburgh. I am struck that the description you have offered us is of a study commissioned by Congress in which you have not characterized--you have not described a persistent set of contacts with your Congressional mentors and in which, if I were a suspicious-minded Senator, I might have concluded you sold out to the enemy and made your deals before you got to me. I am familiar with this kind of working with people who are the subject of studies. Twelve years at the Rand Corporation makes one familiar with that process. But particularly where one is working across two constitutional branches of Government, I wonder if you could not say something more about how much your Congressional clients concerned themselves with whether you are not, for example, putting out premature signals as to which direction the academies should be moving in terms of reform; or conversely, do you want to simply tell us that in a situation like that, the answers are so obvious that the moment concern is expressed, everyone knows which direction to move in?

MR. HARPER:

You have asked a number of questions. I'll take the easiest one and leave the rest of them to Chuck. We in fact did have contact throughout this job with staff people from the various Congressional offices. I participated in meetings with Senator Bayh's and

Senator Proxmire's staff people. Also, as tends to happen in any organization, you get people who come to have specialized functions. On this job, we had someone who was our point of contact with the Congressional people. Our concern was with getting a study done that was as right as it could be from a scientific or methodological point of view because I think both he and I, as well as other people in the organization, saw that the study had great danger for us. There are a group of people up on the Hill who would have liked nothing better than to really challenge whether the service academies ought to exist. There is another group of people that would think that is the worst kind of heresy you could perpetrate. The only way we could win on that one was having something that we could at least defend from a methodological point of view. That was our concern.

MR. THOMPSON:

In terms of setting the direction for change, I think just the fact the Congress was questioning the high and increasing attrition pretty well laid the pressure for where the academy should be directing their attention. I don't think also that our being involved in the academies and with DOD throughout the study had any bias, so to speak, on the results. It was our feeling that it was the best way to go to try to get maximum utilization of what we had to offer when we got through.

PARTICIPANT:

John McGruder from the Department of Transportation. I am curious about whether you looked into historical periods, because it would seem to me that in the mid '70's, with Vietnam over, that you might have found somewhat the same kind of situation that existed after World War II, or after World War I, when it really was not unusual to have a higher attrition in the academies. At least it

would be my expectation that that would be the case. To what extent was this a reason or did you look at it at all? You didn't mention that.

Secondly, your study was obviously a tremendous effort and yet it seems to me as if the principal finding that you came up with was that those who aren't very interested in the program to begin with are much more likely to drop out. And I guess I wasn't too surprised by that finding. Can you help me out a little bit because I am not really amazed by that?

MR. THOMPSON:

Let me deal with the last one first, and John can take the first one last. The commitment one is just the example I used. There are three pages of conclusions and recommendations in the report which deal with all aspects of the academy. In terms of the commitment conclusion, it may seem obvious to you and it does to me; yet, given the program at the academies and the way in which academy people talked about those programs indicated to us that it wasn't as obvious to them.

MR. HARPER:

You see their assumption had been that when the students walked in the door, they were committed. The question was, let's find those who are the most committed and keep them. What we were suggesting was a major policy kind of change in the sense that the emphasis in the program wasn't to be just testing, but was also to be motivating. That is, let's not make that assumption that they are committed. Let's try to develop that commitment in them. It's an expensive process to lose these people.

CONTINUED

2 OF 4

Let me deal with the first question. The point was made very strongly to us by the academies that we should look at historical data. There turned out to be only two academies that had that kind of data, and they were West Point and the Naval Academy. Their attrition after World War I and World War II was quite high. It was low during the depression and that sort of thing. Our major argument for not using it was that there simply wasn't sufficient data to go back and say that there were enough conditions in those periods in time which were similar to the conditions here with only one variable changing--that is, end of war, not end of war. Also if I recall correctly, the West Point and Naval Academy figures didn't jibe. They were conflicting. The Naval Academy's figures showed a steady attrition rate after the wars--steady in terms of what it had been during the war and before the war. West Point was very anxious for us to use those figures, and the Naval Academy wasn't; so we didn't figure they were relevant because of the kinds of arguments that I mentioned.

PARTICIPANT:

I'm Tom Richardson, Department of Commerce. How in fact do you know that the people that are coming out of the system are really the best officers? I would gather that the people who rate them as good mean, they are responsible, they do what they are told, et cetera, et cetera. In fact, given the national interest, perhaps the ones that are dropping out are the best candidates and the ones that are staying in are not. It seems to me that that is an important question to look at, too.

MR. THOMPSON:

No question about it. The problem is we only had two years to do the study. We had to cut off a small piece. The effectiveness of the academies' programs is not only a function, we recognize this,

of the people that go through but also the performance of the graduates. We just didn't have the time or resources to look into that.

MR. RICHARDSON:

In fact you don't know about that?

MR. HARPER:

We present data--what you have in xerox is the main report. There are three appendices that were attached to that. These are printed on both sides of the page so there is a lot of reading matter here. The Appendix B, where we synthesize the academies' studies, presents a lot of information about what the dropouts are like as compared to those who stayed. I suspect you can go through there and form some judgement as to whether they are losing the kinds of people that they should keep. It's a process of socialization. To some extent, at least our review of the studies indicates those they lost at the very beginning are the ones they produce at the end. In other words, they lost leaders right at the beginning because they are not going to socialize to that kind of a system. They are already leaders. Then they go through and mold the others into leaders.

MR. THOMPSON:

I think another important point that came out of our review of the studies was that, as John mentioned earlier, most of the studies we looked at were directed towards trying to control the attrition through controlling selection. There was a very strong reluctance to examine the environment. Our concern there was that over a long period of time, if you began to use the graduates as the criteria by which you selected new students, and then create a cycle, you begin to narrow the diversity of the people that come in to the point where eventually you have one type of person coming out. We weren't convinced this was in the best public interest.

MR. ABERT:

You mentioned that you looked at the characteristics of the students when they came in. I was looking at Barron's 1976 the other day, and I was surprised to see essentially how competitive the military academies are in terms of their ability to attract or at least accept students with the highest academic standards. I would say that across the board in the military academies you are talking about the 97th percentile on Student Aptitude Tests with the Naval Academy a little higher than that. What do you find about retention correlated against high, low, medium SAT scores?

MR. HARPER:

In some academies and in some time frames, there is a positive correlation between measures of academic ability and attrition. When I say positive, what I am talking about are the coefficients that are very small and significant because the Ns are so large. We wound up concluding that while there is some relationship, even those who drop out are of very high ability.

MR. THOMPSON:

Just to give you an example of what we are talking about in terms of the quality of the incoming graduates, in the Air Force Academy for the SAT math score, the average was around 660. The national average was around 460. That is pretty consistent, except for the Merchant Marine Academy, in terms of quality.

PARTICIPANT:

Jim Robinson, Department of Labor. I thought the most interesting thing in your study was something you glossed over very quickly which is that one morning, the academies announced to you that there would be no study or that they wouldn't cooperate with it. Isn't perhaps the most interesting finding in your whole study that there is a basic

question of who is in charge of the military and that the military could sit there and say, "That's it. We are not going to cooperate with your study. Good-bye. Your project is over."

I mean perhaps that is off the point of the evaluation study proper, but certainly the whole idea of your study is that Congress had some right to go in there and ask the military some questions; yet at least some of the military people questioned your authority to even be there and decided it was up to them whether they wanted to cooperate, rather than the other way around under the Constitution.

MR. THOMPSON:

That happens a lot. It's not just DOD. We get questioned a lot about why we are there and we get a lot of flack. I think in the end, we normally get in, so we are used to it.

PARTICIPANT:

Walter Bergman, IRS. One thing I am really interested in, in addition to what you have brought up (obviously, I don't know what is in all those volumes). Did you in any way impact on the selection system? Did you have feedback into the selection process itself?

MR. HARPER:

We discussed in our report the question of whether you could adequately present a picture of the academy to someone who hadn't been there. We weren't talking about changing the selection procedures so much as we were talking about giving adequate information to people about what the academy was like so they could select themselves out before they were nominated or appointed. Our feeling was that there needed to be more of that because the research is quite good in that area to indicate this is a good way to go about bringing people into the organization. But we also talked about perhaps not being able

to do that adequately with the kind of life you are talking about at the academy so what they need to do is identify early in that first summer the people whose commitment is wavering. What we saw in the area of traditionally collected selection variables were such weak relationships that I don't think we would have wanted to make any recommendation anyway. For the most part because they have been selecting top level people, there is no variance. There is very little variance on these measures. They have already been preselected, so you don't find the kind of correlations you normally do. It is hard to develop tests to measure commitment prior to entry.

MR. THOMPSON:

I think also we felt that the academies were doing more than enough research on selection, and in fact our recommendation in the report is that they provide a little more balance in their research and start examining their environment a little bit more in terms of its contribution to attrition.

MR. GRANDY:

Thank you very much, Chuck and John. We'll adjourn our program at this point, pick it up tomorrow morning with the other research papers.

BANQUET ADDRESS NOVEMBER 17, 1976

A CONGRESSIONAL VIEW OF PROGRAM EVALUATION

DONALD ELISBURG, Staff Director and General Counsel
U.S. Committee on Labor and Public Welfare

MR. GRANDY:

As you know, our guest speaker tonight is Donald Elisburg. Mr. Elisburg had a distinguished career in the Department of Labor for quite a few years prior to assuming his position with the Committee on Labor and Public Welfare in the Senate where he is currently the General Counsel. He has his juris doctorate from the University of Chicago and served in the Department of Labor in a variety of positions from 1963 up until 1970 when he joined the Senate staff. He has, I think, a keen appreciation and knowledge of evaluation problems. From our discussion during dinner, I know he has some interesting views. His comments concerning the perspective of the Congress on evaluation is likely to be very helpful to us and thought-provoking. We do appreciate his being able to take time from a busy schedule to address us. Mr. Elisburg.

MR. ELISBURG:

Thank you. As many of you may realize, speakers from the Congress almost always begin with a certain amount of disclaimer. Despite the four or five thousand professional staff people who work for the Congress and the Senate, there are only 535 elected representatives; and when you are employed by a committee, you are responsible to the Chairman (in this case, Senator Williams of New Jersey). I always remark that he is free to disavow anything I have to say on any subject. I'll give you the best views I can, but they are my own and I hope you don't take them as necessarily attributable to the elected officials.

In thinking about the process of evaluation, and the study which was presented to you this afternoon, it occurred to me at dinner that the Congress is really engaged in a tremendous amount of evaluation through its arm of the General Accounting Office, but we don't normally think about it in terms of evaluation. We think about it as having GAO do a study, or GAO do an investigation, or something breaks loose in the newspaper and you say, "Oh, boy, we better do something about it," so they send the GAO in to take a look. That really is a rather extensive investigative arm and, therefore, to some degree, an evaluative arm of Congress. Lest anybody think this is not a significant kind of career that those of you in the business have embarked on, this afternoon there came across my desk one of those documents that you never look at. But because it was a very nice package and had a little short note clipped to it, I decided I would at least take a look at it before I put it in my outbox. The title of this nice book is, "Recurring Reports to the Congress: a Directory. 1976 Congressional Resource Book Series."

The note says that this comes from the GAO and a copy is enclosed for your use. The third volume, Federal Program Evaluations, will be distributed in December. This particular document is, I guess, a list of various kinds of reports that come to the Congress for the umteen thousand statutes that exist. The point is that this source book and Volume I which came out about six weeks ago, listed the various kinds of data collection processes that various agencies use. It was a red book, compiling lists of things that the Government is doing and that you are all doing, either as members of the Federal establishment or engaged in some relation to it. My point is that this kind of compilation never existed before. People in Congress probably have no idea of what they have fostered over the years, and perhaps after a couple years of putting this book out they will be sorry they ever got into it.

But I think it's clear we are in some new arenas of doing business with each other. Consequently I really appreciate the invitation to address this symposium.

There are always innumerable conferences going on in Washington. When you think about it, though, I dare say few, if any, are as important in substance as this one which examines how Federal agencies utilize program evaluation. The subject requires intense consideration if the Executive Branch is going to maximize its administrative responsibilities in implementing programs fostered and enacted by the Congress.

Personally I am pleased that this part of the symposium in process includes the Congressional view of evaluation as well. Hopefully the participation of those of us who are connected in some way with the Congress will contribute to the success of your very timely and essential meeting.

The Congress differs markedly from the administrative agencies with which many of you are associated. It is a body responsive to the wishes of multiple, sometimes conflicting, sometimes shifting constituencies. Almost all Governments have an Executive Branch. Our nation is one of the few in the world which entrusts its lawmaking to an independent, periodically elected representative body. And that body functions in a milieu which, by its very nature, is heavily politicized.

By politicized, I mean that the Congress listens carefully and continuously to its broad array of constituencies. I think that this listening is the Members' first and most basic source of evaluation: it is a very finely tuned antenna. Sometimes the listening is carried out scientifically through the use of sample surveys. Sometimes it is carried out intuitively, as when rumblings and

grumblings, wishes and preferences are brought to awareness by the delegations to the offices of the Senators, Representatives or Committees. The unrelenting deluge of mail and the representatives of special interest groups bringing their clients' wishes and complaints to Congressional attention are two additional sources of evaluation playing a role in the assessment of policies and programs.

This is probably the method of evaluation prescribed or implied in the Constitution. It is the means by which our system of Government has worked for two hundred years. It is not a perfect system as everyone knows, but it has been, on the whole, a successful device to balance overwhelming societal concerns with individual liberty and rights.

In spite of cyclical praise or scorn, the Congress has maintained, as its primary means of evaluation, the legislative judgment for which it is accountable to the electorate. This basic fact alone conditions the way in which more systematic, scientific program evaluation is viewed by the Congress.

While growing recognition of professionally-based, expertly-conducted program evaluation has been evident in recent years in the Congress, legislators and their staffs view this important secondary evaluation supplement within the political framework Constitutionally required of them.

Systematic impact assessment of Government policies and programs has been accorded increasing acceptance by Congress. However, the products of such assessments are looked upon as tools with which to shape the essential substance of programs attracting a following or an opposition among the constituencies having an interest in them, including that amorphous electorate whose opinions are often made known only on Election Day.

It cannot be stressed too strenuously that scientific program evaluation is itself evaluated by the Congress in terms of its utility to promote the effectiveness and precision of legislative judgments in a political milieu.

In recent years, program evaluation has been made a requirement of many policies and programs enacted into law. In some cases, this requirement has taken the form of directives to Cabinet Officials to set aside and allocate a fixed proportion of funds to evaluate a selected program. In other cases, impact statements have been required. Impact statements can utilize program evaluation together with other research devices designed to provide assessments of net impact. Both forms of Congressionally-mandated evaluation have the same purpose: to delegate to the Executive Branch a duty to determine what if anything happened as a consequence of the policies or programs tagged for special review. This may be another form of cop-out, perhaps by the Congress, but it is a way of putting the burden on the Executive Branch to do the work.

More recently, Congress has asked directly whether or not our policies and programs are cost-effective--whether we as the public are getting "the right authorized impact of the legislatively appropriated dollar," and whether the nation's economic interests and social well-being have in fact been promoted, especially by human resource efforts.

Much of this budget-related interest emanates from the provisions and procedures of the Congressional Budget and Impoundments Control Act of 1974. Many Senators and Representatives favor the Act, in part because it provides a budgetary window into the inner workings of programs. That window is made possible by fiscal analysis and fiscal priority-setting. I would venture to say that many Senators

and Representatives are not totally pleased with the Budget Act because the scientifically-developed new procedures--that require Authorizing Committees to look at specific dollar amounts and that require an overall picture of what is being appropriated and what isn't--have raised very serious questions, particularly in the social arena, like, How do you keep from being shortchanged because you don't come up with the right numbers on the computer?

More recently, Congress has begun to consider in the formulation of the Sunset Bill, steps which would institutionalize program evaluation and review at the heart of Congressional decision-making. In the Senate, for example, the Government Spending and Economy Act of 1976--that was the Sunset Bill--proposed that programs be terminated on a mandatory basis every five years and reauthorized only after a close-scrutiny program review. Fortunately, the bill was not acted upon, but the idea has attracted a following in Congress. While legislation of this nature has many features, the degree of dependence on program review techniques would be tremendous. Were Sunset Legislation enacted in some form, recognition of program evaluation as a secondary means of Congressional decision-making would have attained an enhanced status. I think it would be nicknamed the Evaluators' Full Employment Act. It would also have been accorded grave responsibility as an instrument of public trust.

But even if such developments were to occur, would the public trust indeed be well-placed? What would the Congress be buying?

Many experienced Congressmen and their staffs are concerned that the Congress will become dependent upon a program evaluation establishment--valuable in concept, but unproven in product. Opponents of the systematic use of program evaluation point out that such research

is an art form of marginal reliability and that reliance upon such an art form is in itself more in the nature of folk medicine than of science.

The issues cited as the source of such suspicion are commonplace:

- that the assumed posture of objectivity among program evaluators often masks subtle but important biases and hidden agendas;
- that the questions set for discovery, if published at all for client consideration, have predetermined answers;
- that the procedures utilized frequently neglect the most important variables often included in initial designs and later dropped because of difficulty in research management or unexpected costs;
- that there persists an inability or unwillingness to merge the contours of various impact evaluation studies so that common patterns of findings can be codified and differences in findings highlighted;
- that interpretations of findings are cast in terms far in excess of their value and far overstated to listening audiences; and
- that the conduct and packaging of evaluative research supports first the publication interests of the investigators and too often relegates the needs of clients and sponsors to second place.

Whether or not these assertions can be supported substantially, the doubts exist. Program evaluation experts will point out that the Congress has its own peculiarities, biases and statements which lose support when subjected to rigorous analysis. But the Congress bears the accountability of the electoral process in setting forth its assertions into law, overseeing its handiwork, and supporting its decisions from the Federal Treasury. Obviously, the task before us is to look beyond the concerns (while keeping them in mind) in order to explore some principles which would enhance the utilization of program evaluation by the Congress. The task requiring attention

is to develop program evaluation standards and approaches which will notably assist the Congress in its accountability for public policy.

If program evaluation is to become truly useful to the Congress, those conducting research as the agents of the elected officials should consider three principles in adapting their works to the needs of the Federal Legislature.

The first principle, and perhaps the most difficult to achieve, is that program evaluation must be preceded by policy analysis and mission analysis. Policy analysis in turn calls for a rigorous study of the substance of the policies giving rise to programs. Policy analysis calls for the consideration of the goals enunciated during the formulation of policies and programs. Policy analysis requires attention to drift and shift between policy as legislatively mandated, and policy as executively implemented. Policy analysis requires careful attention to the process, the actors, the subtle differences which result in a policy product.

Mission analysis is the description and explanation of whether a program adheres to the objectives set forth in the policy. The fundamental question of concern to the Congress is whether a program carries out the mission established for it in the policy. It is to that issue that constituent concerns are addressed as well.

I would stress here that policy and mission analysis require as much research skill and time as any other element demanding the attention of the program evaluator. Policy analysis requires case study techniques; selective use of surveys; employment of content analysis of documents; and utilization of journalistic and investigative techniques which employ accepted standards of corroboration. It also means you have to be able to write clearly.

The foregoing implies, of course, that the researcher can generate the trust necessary to conduct an adequate policy analysis as a preliminary step to informing the Congress about the impact of programs. But that trust is essential inasmuch as the questions central to program evaluation are likely to be derived from policy and mission analysis.

I would also add here as an important factor that most members of Congress and their staff have been trained with legal concepts and investigative techniques. It is not surprising then that they frequently regard the standard of evidence utilized by many program evaluators as inadequate. When one reads through program evaluation reports and is struck by the large number of tables pronouncing this test as statistically significant and that test as unassailable evidence of a particular program impact and one reads on further to find that conclusions have been drawn entirely from aggregating such statistical inferences of proof, it is not surprising that the clear and convincing evidence standards, or the preponderance of the weight of the evidence standard used in legal thinking seems, by contrast, far more reliable.

In short, persons connected with the legislative process are not likely to be convinced that large numbers of associations of variables prove a point. Common sense requires complex situations to be judged with all available evidence--both the context of the situation and the specific variables artificially isolated for examination--before conclusions can be made. That is scientific jargon for saying that you must do a careful job. Program evaluation and policy analysis, in particular, will be judged by the Congress according to a standard of evidence not usually advanced in the program evaluation with which many of you may be involved.

A second principle to assist the adaptation of program evaluation to legislative activities is that the evaluator must understand where the Congress will find evaluation useful. While evaluation studies may be useful in the formulation of bills, program evaluation is most relevant to Congressional oversight. Congressional oversight is a shorthand term we all use for what we do, with a broad license to do anything, after a statute has been passed. Congressional oversight is the means by which Congress accounts for the policies and programs it authorizes and appropriates. The common techniques utilized in Congressional oversight include investigations; hearings; site visits; audits; analyses of special and recurring reports required by statutes; meetings and meetings and meetings to consider the impact of appropriations of funds for program support; procedures to consider formulation of the Federal budget under the provisions of the Budget Act, and so on and so on. Obviously program evaluation could have a strong role to play in some of these activities, a lesser role in others. The important point is that an understanding of the conduct of oversight is itself important. Familiarization with the techniques utilized, procedures employed and the settings for oversight activity cannot be substituted.

Finally, I would suggest that attention be given to the way in which program evaluation studies are interpreted, presented and packaged. Congress, I am sure, is acutely aware that various constituencies in a political milieu may be activated in favor of or opposed to a program by the expert character of an evaluation report. I might also add that the Congress or individual Senators or Congressmen may well be influenced by whether you can relate the five years of your evaluation study in the 15 minutes that you have at a public hearing, that is, how well you can do it, how well you can synthesize and set forth, while you are sitting there on a TV camera, the essentials of what you have been trying to do.

One is also acutely aware that public hopes ride high on programs finally forged from Congressional actions. Hard-won advances, particularly in the human resources field, may suffer permanent, unwarranted damage if the evaluation and interpretations are unjustifiably sweeping, if packaging is conducive to sensationalism in the public media and if presentation does not relate to the concepts or procedures conventionally employed by the Congress. Hopefully, this foregoing litany will provide some basis for your discussions tomorrow as the business of the symposium proceeds. I think the Legislative Branch has an important stake in program evaluation as it goes about making and shaping the public policy with which we are all going to live. The prospects for Congressional utilization of program evaluation are very great. In our own Labor and Public Welfare Committee, we have begun for the first time in its history to institutionalize some of the evaluation ideas; and that primarily means to appoint relatively permanent staff to think about it. That is a big step. It's a big step in a fairly tight-budgeted operation, where you have relatively small numbers of people, to assign someone to start thinking about the evaluation of programs and something resembling a systematized oversight.

The evaluation research and the people who conduct it--that is, all of you--may very well become a very important augmentation to the fundamental framework of legislative decision-making. You may not all welcome the prospect, but I think it's more than just around the corner. It's true because of the Budget Act and many of the other possibilities, the Sunset Act, for example, that have evolved around the Congress, and the fact that the Congress is now dealing with a budget of some $400 million and some odd a year, really a billion a year. Evaluation is really a kind of program technique that is not new, but it is

something that is going to increase its respectability; and consequently, I think it is going to be an important adjunct to the legislative process. Thank you.

PARTICIPANT:

I liked your speech very much. Congress is my first interest, and I was pulled into criminal justice for want of a job. We have been talking about the need to specify objectives, and Congress often passes acts like the Crime Control Act that says, "Reduce crime and improve efficiency." The Act itself has to specify the objectives of that Act, and how do you measure that? Is it reasonable to expect that Congress might specify objectives very clearly--that the objective of an act could be very specifically stated in the act itself?

I give as an example the Speedy Trial Act of 1974, where Congress not only specified the objectives, and showed how to figure out whether a speedy trial is achieved, but wrote the evaluation design into the Act itself. It has been done at least in that one case. The point is that agency administrators never specify objectives. Can Congress do it?

MR. ELISBURG:

I understand the point. The legislative process does not lend itself to regulation writing. By and large when the Congress has gotten into writing in detail the specifications of how it wants something carried out, it either gets into trouble or the events of time pass it by; and you have to relegislate. With respect to the question of being able to spell out the policy, however, almost every major piece of legislation has a findings and purpose section which can go on ad infinitum trying to spell it out. I would recommend to anyone dealing with a serious legislative enactment, for example, a major program, that you look not just at the words in the statute, but

that you look at the legislative history and the committee reports, I think you will find from these sources a more detailed program lay-out of what it is the Congress intended and wants done with these programs.

PARTICIPANT:

I have a two-part question. The first part deals with your statement concerning policy evaluation. I'd say that all during the day, there has been some question about what kind of focus we should have in our evaluation. Some suggestions were that evaluation would be better if it were narrowed to doing what our boss wants to see. Others, more expansive, wanted to do what everybody wants to see. What would your views be on that?

My second point has to do with the acceptance by the Congressional Representatives of this information. Friends of mine in the Congressional Budget Office say that, in fact, they feel that their activities are viewed by many of the Representatives as constraining. The fact that they come up with facts means the decisions that the Representatives can make are somewhat weakened in certain lights. I would see evaluation providing the same kind of data which would be equally constraining. How do you feel about that?

MR. ELISBURG:

As to the first part, I would view the question of how an evaluation should be done from the standpoint of whether I was the boss or everybody else. I think the problem is really of defining the policy. You have to really take the time to understand in a legislative context what it is that the Congress had in mind, what the objectives were, and how those objectives have been met? What was it that the Congress was trying to set forth? Otherwise you might just as well be evaluating apples when Congress is talking about oranges.

The second part as, to whether anybody is going to feel constrained, is really a question of growth and development of the institution with which you are dealing. Fifteen years ago, the Senate Labor and Public Welfare Committee had a dozen employees and very few statutes within its jurisdiction. The Senate Labor and Public Welfare Committee now has in excess of 125 employees, which is not necessarily large in terms of a Government agency, but it is responsible for reviewing programs which represent in excess of $40 billion a year. There are literally hundreds of them. When you are talking about legislatures which have to deal with that kind of fantastic growth in legislative programs, newer techniques will have to be used. For the first time in the history of the Senate, really, we are getting a computer capability that most Federal agencies had 15 years ago and most of private industry had 20 years ago. It's a growth process. There is a realization and understanding that these techniques are going to have to be used, constraining or not. So to that extent, Congress, like a lot of other groups, is being dragged kicking and screaming into the 20th century. Thank you.

IV. THE HIGH IMPACT ANTI-CRIME PROGRAM: A PROCESS EVALUATION

ELEANOR CHELIMSKY, Department Head,
Program Evaluation Department
The METREK Division of The MITRE Corporation

MR. GRANDY:

As you are aware, yesterday we fell behind in our schedule. I am not too concerned about that. I think the relaxed and candid interchange of ideas and information is worth it, and I admire your perseverance and stamina in sticking with us. We have two papers left from yesterday afternoon's session which we will begin with this morning. It is our plan to delay our luncheon one hour, so we will go to lunch at just after 12:45 instead of 11:45. This will somewhat shortchange our afternoon working panels. We will try to make up time there later in the afternoon.

This morning our program will start with a presentation of another research paper by Eleanor Chelimsky who is Head of Program Evaluation at the METREK Division of MITRE. Her paper concerns an evaluation conducted of LEAA's High Impact Anti-Crime program for the National Institute of Law Enforcement and Criminal Justice.

Ms. Chelimsky is an economist by training, served as a statistical analyst at the U.S. Mission to NATO, and, since 1970, has held a variety of research positions at the MITRE Corporation. Most recently, she has directed policy analysis and program assessment in the areas of health, welfare and criminal justice. She presently heads up our program evaluation department. Eleanor.

MS. CHELIMSKY:

Thank you. Well, as Cork has just said, I am going to talk to you today about the national evaluation of the High-Impact Anti-Crime program which MITRE performed between July of 1972 and December of 1975. There is a summary of this evaluation on the table outside, and it may be useful to look at it because I know that in the short time I have, I am not going to be able to do more than give you a very broad-brush and generalized account both of the evaluation and of the findings.

Before examining them though, I'd like to look just a little at the program itself and at the origins of the program, not because their bureaucratic and political aspects are especially unusual, but--in the sense that Jim Stockdill was talking about yesterday[21]--because they help to explain the program and some of its peculiarities--its ambitiousness, for example, and its unusual complexities--and because they also say something about the agency needs which drove our evaluation.

When you go back to the crime control context of 1971, perhaps the first think you need to remember is that the Nixon Administration had been in office for about three years, and that the 1968 campaign had focused very heavily on crime as a political issue. Although the Safe Streets Act had created LEAA in 1968, the crime problem had not abated by 1971, as many people pointed out yesterday. Another election was coming up in 1972 and it seemed to be a propitious time for a major new anti-crime initiative. Also, by 1971, LEAA seemed to be coming out of the turmoil which had marked it since its creation,

[21]See pages 129-132 above.

turmoil due, at least in part to the troika organization that Congress had imposed on it. So in 1971, there was not only an Administration need for a big, visible, ambitious anti-crime program, there also seemed to be an agency capability to mount such a program.

Another factor which explains the ambitiousness of the Impact program was the still optimistic, gung-ho climate of 1971. It seems a little strange to remember it now, but it was common then to hear people saying things like, "If we can send a man to the moon, we can... fix the economy, or cure cancer, or turn the corner on crime and drugs," or a hundred other good things.

It is true that researchers were not quite so optimistic at that time, after the poverty programs of the '60's, but their caution doesn't seem to have penetrated the upper reaches of administration where programs are born and made. At least, not then. In fact, there was real optimism about the potential of a concentrated thrust for "doing something" about crime.

As for the complexities of the program, some of these can, I think, be traced to policy issues that were confronting LEAA at that time. Many of them had to do with the fact, as Dick Linster said yesterday, that LEAA is basically a block grant program. LEAA is, and must be, concerned with the problem of working with states and localities. Some of the issues surfacing in 1971 concerned questions like: How can Federal leadership be made acceptable to states and localities in an area where they had had undisputed primacy three years earlier? How do you apply Federal resources to local crime problems so that local people have a dominant voice in deciding how the money gets spent and at the same time insure that the money is not misapplied or misappropriated? How do you go even further and insure not only

that it's not misapplied or misappropriated, but that it's effectively spent? What kinds of analytical capabilities do you need to build in at the local level in order to do that?

Even if you can get states and localities to accept Federal leadership, how do you make that leadership effective in terms of results, given that there is something of a gap between research capabilities at the Federal level and research capabilities at the state and local level, and an even greater gap between researchers in general and criminal justice practitioners who need to use and apply their research? How do you insure that Federal research can be disseminated, and also understood and used by criminal justice practitioners at the state and local level? That's a pretty tough question.

How do you overcome the reluctance of independent agencies to coordinate their efforts when very often it seems to them that they have, in Sam Seeman's terms yesterday,[22] little practical incentive to coordinate, and a great many incentives to avoid coordination? How do you get them to include the public in their planning and program processes when again, there are real disincentives to do so, despite all the studies which have shown that it's important for the success of federal programs to involve the public in their planning and operation?

All of these were major policy questions for LEAA in 1971 and they all of them found their way into the Impact program.

Still another source of complexity in the program was the professional criticism that had been heaped on LEAA in 1971 in the Task Force report. There were four general areas of criticism raised

---

[22] See 21 above.

in the report. The first one was that state and local recipients of LEAA block grants were squandering a great deal of money, and that LEAA had failed to perform an adequate fiscal monitoring job.

The second area of criticism was that too much money was going into police hardware.

The third area was that not enough money was going to corrections and specifically, to rehabilitation programs.

Fourthly, the Congressional report said that evaluation standards hadn't been built into LEAA programs so that it was difficult to judge their effectiveness.

The final source of complexity in the Impact program which I want to mention here is just precisely this question of evaluation itself. It seemed to many people at LEAA that evaluation could be a very promising tool not only for discovering whether programs work or not, but also for doing what LEAA wanted to do in the area of upgrading state and local analytical capabilities. But the fact was that, in 1971, no one really knew how to do that. There were not many social program evaluators around in 1971; there was no great pool of expertise to draw on.

In sum, the context that I have been looking at here points to the emergence of a very special kind of anti-crime program: big, visible and ambitious; highly complex; focused on corrections rather than on the police; locally run but financially unassailable; and containing a major effort to upgrade system and research capabilities at the state and local levels.

Vice President Agnew launched the program in January of 1972, very very visibly. The program was to be sizable: $160 million over two fiscal years to aid crime control in eight U.S. cities (Atlanta, Baltimore, Cleveland, Dallas, Denver, Newark, Portland and St. Louis). These cities were asked to have their programs operational within six months--that is, "on-the-street" and working by July of 1972.

To understand what this meant in terms of the enormity of the local implementation problem, you need to look a little bit at the criminal justice budgets of these cities before they go this $20 million increment; it meant different things to different cities. For Baltimore, with an annual criminal justice expenditure of $72 million, a $10 million increase did not seem so very indigestible. For Atlanta, on the other hand, with a total expenditure of $15 million, city efforts to absorb the Federal funds resembled those of a cobra trying to swallow a piano. But for all of the cities, the questions of how that $10 million should be spent, and what mechanisms could be found by which to spend it, were major problems.

The modus operandi of the program was New Federalism. Briefly put, this is the idea that local priorities ought to be set by local people. The cities were told that they could develop their own programs, run them and evaluate them according to their own criteria. In this way it seems that LEAA was avoiding coming to grips with the Federal leadership question and was instead proposing an equal, Federal-local partnership. The local control that is implied by New Federalism, however, was going to be tempered and corrected by a very tight fiscal and program review that would be done by LEAA's state planning agencies and regional officers.

The most important means of upgrading system and research capabilities at the local level would be the crime analysis team, a group

of researchers and criminal justice practitioners who were to be established in each city. Their function was, first, to supervise the performance of the highly complicated Crime-Oriented Planning, Implementation and Evaluation process (a mouthful of jargon which we call the COPIE-cycle, to shorten it). Second, they were to do what they could to improve agency coordination; and finally, they were expected to involve the community, to the degree possible, in the workings of criminal justice plans and programs.

The COPIE-cycle was clearly a very complex operation. The cities were being asked to collect a great deal of data (much of which was not in existence) about their crime problems. They were supposed to look at local data on victims, offenders and crime settings to get some real sense, based on the data, of what their problems actually were. Then they were supposed to rank their problems, achieve some consensus on their priorities among the various agencies of the criminal justice system, develop programs to address their crime problems in some reasonable way, build evaluation components into their programs, and finally, evaluate them.

The program did not target law enforcement alone, but rather a comprehensive, across the board, anti-crime focus which addressed the Congressional criticism about police hardware. The program would specifically encourage and emphasize corrections programs through a fiscal incentive: cities only had to provide 10 percent matching funds for corrections projects (as opposed to a 25 percent local match for other kinds of efforts).

At least two conflicts in the program are immediately apparent. The first is that the cities were told that the program would be theirs to run, yet the emphasis on corrections which is a state function, signified that this could not really be the case.

The second conflict is that the programs were expected to be operational in six months; yet is is hard to see how the cities could get through the COPE-cycle and also have their programs implemented in time, especially since most of the teams that were supposed to supervise the cycle weren't yet hired, or in residence in the city.

In practice, it turned out that for the cycle to be performed in a reasonable way, it took about sixteen months.

The major objectives of the program were six (see Figure 7 below), and they are typical of the objectives of most broad-aim, action programs. That is, they are not operationally defined, and they fit to a "T" Bob Hemmes' description yesterday of vague, virtuous and desirable goals like "support civilization."[23]

The first objective was to reduce crime (that is, decrease stranger-to-stranger street crime and burglary) and the stranger-to-stranger street crimes targeted were murder, aggravated assault, robbery, and forcible rape. It was hoped that these crimes, as well as burglary,could be reduced by 5 percent in two years and 20 percent in five years. Now this objective may seem more specific than the others, more quantified, but basically it was meaningless because the cities were going to develop their own programs. They hadn't yet even started to think about them when the objectives were announced, and they had a choice of project options which could affect crime rates differentially, unmeasurably, or not at all, so that there was no way to determine in advance what crime decreases might logically be expected from a program still to take shape.

Planning for the Impact program was forcibly curtailed by the great rush to speedy implementation. There were, perhaps, three months of program planning performed in all, but almost no evaluation planning at the national level, except to decide that there would be three levels of program evaluation--city-level, national-level, and a macro or global level.

---

[23]See page 71 above.

FIGURE 7

PROGRAM OBJECTIVES
The High-impact Anti-crime Program

- Reduce crime: Decrease stranger-to-stranger "street crime" and burglary by 5% in 2 years and 20% in 5 years
- Demonstrate the copie-cycle and test the crime analysis team
- Acquire new knowledge about crime
- Improve coordination among criminal justice agencies
- Increase community involvement
- Institutionalize innovative, effective projects

city-level evaluations were expected to produce findings of effectiveness for all the projects that would be implemented in the cities. (It was mandated at the start that every project would be evaluated.)

The national evaluation was supposed to look at program activities and processes within and across the eight cities, using data generated by the cities as building blocks.

The macroevaluation would examine the anti-crime effectiveness of the program using victimization surveys. This evaluation was intended to be performed by the Statistics Division of LEAA in combination with the Bureau of the Census.

We, MITRE, contracted to do the national evaluation in July, 1972--about six months after the program began, and worked closely with the National Institute to develop an evaluation plan. We knew we wouldn't be looking at overall anti-crime effectiveness because the global evaluation was going to do that. And we knew that we couldn't very well impose an experimental design on this free-form, New Federalist program that was going to be totally different in each city and didn't allow the possibility of central data collection. (All of our data was to come from the city-level evaluations.)

We felt there were a great many process questions to answer and we tried to identify, among the multitude of possible inquiries, what LEAA and the National Institute were really hoping to find out from the program. So, moving toward a process evaluation which would ask the question, "what happened?" rather than, "did it work?", we began to examine the researchability of questions like, How feasible, in fact, is the OPIE-cycle at the city level? If it is feasible, if the cities

arrive at performing it, does it allow, at the national level, some ability to determine whether city programs are effective or not? Does it improve research capabilities at the local level? How useful is the crime analysis team? Is it actually possible to do something about agency coordination? How likely is it that the team can be successful in getting people in high-crime, inner city communities involved and concerned with criminal justice? How viable is New Federalism as a program philosophy? What happens when the time comes to get city compliance with program requirements (when we, MITRE, need to ensure that city data has been collected and evaluations reported so we can do our own evaluation), and there are not teeth in the program with which to do so? How reasonable is it to expect objectivity in city evaluations of their own anti-crime projects?

What kinds of projects do cities generate when the Federal Government gives them $20 million and tells them to do crime analysis? What happens in that process? If they are effective, those programs, do they get institutionalized? Or does the whole thing just fade away when the Federal money goes? What are the lessons we can learn in terms of future programs? Those were the kinds of questions we wanted to look at.

Together with the National Institute, we eventually developed an evaluation plan which contained eight tasks in four general areas. Our major process mechanism was a program history in each city which featured interviews with a great many people during and after their involvement with the program. In those histories we looked at program development; at key actors and their roles; at the ways in which the crime analysis teams functioned, what they were doing to attack some of the problems that they had, where they were succeeding (if they were) and what their techniques were; and we looked at various types of city-state power relationships (much like those John Harper described yesterday)[24] across the criminal justice agency spectrum.

[24]See page 173 above.

To observe the COPIE-cycle, we did in-depth examinations of city-level planning, evaluation planning, implementation and evaluation reporting. We looked at what speeded up implementation, what slowed it, where the bottlenecks were, what the quality was of evaluation planning and evaluation reporting. We did get a great deal of data in those areas, data which now furnish an interesting baseline of what local capabilities were in 1972, in terms of planning and evaluation.

At the beginning, we were hoping to do cross-city studies of commonly-encountered strategies and problems. What we found was that these projects were simply not comparable. In nearly all our fields of effort, what we got basically were case studies. But we did look across the cities and compare these studies, examining areas like drug treatment strategies, police patrol efforts, intensive supervision for juvenile probationers. Again, across the cities, we looked at caseload and trial delay problems in felony courts, which gave us an unhappy familiarity with the recordkeeping systems of some of our courts.

We looked carefully, across the program, for signs of project innovation. Although innovation was not a major objective of the Impact program, everybody was hoping, nonetheless, that there might be some exciting new projects developed despite the difficult analytical constraints of the program. Finally, it turned out that there weren't many, but we did find some, mostly in the area of community-focused projects. We also looked at projects to see if they might be likely candidates for transfer to other places, and we tried to see what could be determined about the probability of institutionalization for many of these projects.

What happened, then, in the Impact program? In all, over three years, the Impact cities implemented 233 action projects, and those projects cost about $140 million in Federal funds. Generally speaking, the program did focus on corrections--in particular, on the juvenile recidivist offender (see Figure 8)

If you look at the Impact program according to the objectives of each of the projects--and this is possible because, with evaluation planning built in, we had a fairly clear record of precisely what was being expected of each one of these projects--you can divide it into three thrusts or foci. There was a straightforward crime reduction focus which essentially involved police programs, street lighting programs, crime prevention programs. Some of these were community based, and some of them were police based, but all of them had as their intention to reduce crime in a particular area. About 31 percent of program funds went to that kind of effort.

Forty-two percent of program funds went to recidivism reduction, which was essentially an effort to treat, find jobs for, counsel, rehabilitate, individual offenders via various correctional or diversional alternatives.

Finally, what we characterized as a focus on improvement in system capability (that is, efforts which tried to increase capabilities through data systems, research, better management, that kind of thing) accounted for 27 percent of the Federal funds.

Thus the Impact program was not essentially a deterrence program (as it has sometimes mistakenly been called) but was rather a comprehensive criminal justice effort with its major emphasis on offender treatment.

FIGURE 8

Dimensions of The High-impact Anti-crime Program in Terms of Emphasis

Project Focus	Percent of Impact Funding (%)	Funding
Crime Reduction	31	$ 44.0 M
Recidivism Reduction	42	58.4
Improvement in System Capability	27	37.6
Total	100%	$140.0 M

Let me just try to summarize now very briefly what our general findings were. Overall, they fall into two gross categories: findings on the objectives, and findings on program management.

The findings on the objectives deal essentially with the COPIE-cycle (that is, the Crime-Oriented Planning, Implementation, and Evaluation cycle which I discussed earlier), with the crime analysis team, with project effectiveness at the city level, and with project institutionalization.

After looking at the various segments of the COPIE-cycle in depth, looking at planning, implementation and the rest in each of the eight cities and across all of them, we found that despite the newness of the concept and the difficulties of implementation, and despite the lack of enforcement mechanisms, all of the eight cities actually did perform this complicated thing. Some of them performed it well (there were four cities that did very creditable jobs) and some of them performed it less well. But we found evidence of quite notable increases in analytical capabilities (new efforts undertaken, new approaches, new products generated), and in research capability, generally, wherever it was performed.

We found evidence that the crime analysis team was effective, but effective under certain circumstances only, quite outside the question of the professional and personal characteristics of the people who were in the teams. It seems that organizational locus was extremely important. When the crime analysis team was in the Mayor's office, or was closely affiliated with it, benefiting from the support and power of the office, it could barter effectively with the various criminal justice agencies; and that was really the essential point in

its ability either to supervise the complicated COPIE-cycle process or to do anything about coordinating agencies. When the team was located elsewhere, it tended to be ignored and to founder.

We also saw that when the team was deprived of the evaluation function, which happened in two cities, it was considerably weakened. Apparently the ability to work closely with agency managers which accompanied the evaluation function, was very important in getting the agencies to accept them. Technical assistance in evaluation was a quid pro quo which could be offered in return for cooperation or coordination. When the teams didn't have that possibility, again they were much less effective.

Four of the crime analysis teams improved agency coordination in their cities. Part of this was due simply to the inauguration of a process whereby staff people from different agencies were obliged to talk to each other on a regular basis. In Cleveland, for example, probation and parole people began working closely together in ways which they had not done before. Eventually, both groups were housed together in the same building. Before Impact, those people didn't speak to each other. There were all kinds of things of that sort that occurred, that were made to happen.

In Denver, a community mechanism was developed which they called the Neighborhoods Task Force. This task force, recruited in the community on a volunteer basis, worked regularly with agencies and the public throughout the whole program. People went out into the community, and it was a little like getting out the vote. They actually got community members involved and to meetings; every month during the program, there were real interchanges among judges, police, people in

all areas of the criminal justice system and the communities they were serving. Some of the meetings were quite heated at times because many people were disturbed by some of the programs which they felt were being foisted on them. These were real interchanges, not lip-service; by the end of the program, new procedures for consulting involved communities before program implementation had developed in Denver.

The COPIE-cycle did permit us to examine project-level effectiveness. We performed secondary analysis, and were able to reinforce city claims of success in reducing crime or recidivism in quite a few instances, accounting for about $35 million of Federal funds. This doesn't mean that the projects we looked at were the only ones which may have been successful. They were, however, the only ones that had evaluations rigorous enough so that we could attempt to validate their claims.

Our inquiries showed that about 43 percent of the projects funded were set to be institutionalized in one form or another. We believe this to be unlikely, based on past performance in similar programs. Obviously, you would have to return to the cities a year or two from now and see what really comes to pass. The final number will probably be closer to 25 percent, or something like that. (Even that would be very good, however, compared to many other Federal programs.) We did find that institutionalization appeared to depend much more on the support of key personnel than it did on whether the project was good or not, which rather threatens the conventional wisdom.

In terms of program management, we found that New Federalism was much more of a hindrance than a help. It isn't even clear that it elicited the local priorities it was supposed to elicit. I think the data analysis did more for developing priorities than did New Federalism because what really happened was that when you could show, in Baltimore, for example, via data, that you had a tremendous

aggravated assault problem, or in Portland, that the problems were really robbery and burglary, it then became difficult for people to take projects off the shelf and say, "we need to do this or that," when there was no data there to support it. Further, New Federalism was a great hindrance in getting cities to do what they had contracted to do, because the philosophy precluded enforcement mechanisms in the program.

We found evidence that the fiscal review was successful. There seem to have been very few dollars that weren't accounted for in the Impact program.

The program review was much less successful, on the other hand. The state planning agencies and the regional offices didn't have the personnel to do the program monitoring that had to be done, and they didn't have the expertise to review the evaluation plans and reports which needed careful review. The program review was also excessively slow and caused a lot of irritation in the cities. We found that technical assistance to the cities, especially in evaluation, was generally lacking, and I guess this was part of the overall evaluation problem of the period. People didn't realize how much technical assistance was needed. We also found that the absence of national evaluation planning was a serious loss to the program because, of course, a great deal more and better information could have been collected if program development had been accompanied by evaluation planning.

Finally, we found significant data problems. These are much too long and complicated to go into here, but I really would like to quickly mention four of them. First, inadequate agency record-keeping, especially in courts and corrections--there were major gaps and inconsistencies in the records which caused serious problems to city evaluators.

Second, difficulty in using UCR data. I guess everybody knows about that, the discretionary problems, the difficulties that are involved there.

Third, there is a lack of standardized data for measuring recidivism. You can say all you want about how terrible the UCR's are, but they exist. They are there. You can, if you want, go to look at your crime-reduction program, see what you are getting and measure your results against the UCR's. There is nothing to measure recidivism outcomes against. This is an important gap; there is a great need for a standardized data base in this area.

Finally, there was the crucial inability in any Impact city (or elsewhere to my knowledge) to trace an offender from his point of entry into the criminal justice system until his return to society. This meant you couldn't really look at what was happening in your programs and at their impacts, except in little segments. This was again a major problem for evaluation.

All of these data problems again reduced the amount of technical information which evaluation at any level could produce in Impact.

We think LEAA has made considerable use of our findings and recommendations. They have been explicitly examined and incorporated into planning for new programs. One of these programs now specifically implements our recommendations for greatly increased technical assistance to localities; for phased program approval which could put teeth in a program in the sense that we didn't have them in Impact; for management information systems to flag operational problems; for increased program monitoring; and finally, for a much amended and improved COPIE-cyle.

I think there are several reasons why our findings were used by LEAA. First, there was great continuity in the person of the program manager, Dick Barnes, Head of the National Evaluation Program at the National Institute, who endured the stresses and strains of Impact from beginning to end. Despite profound and frequent changes in philosophical approach and in personnel at LEAA and at the National Institute, the program was never "lost from view" or "lacking an organizational home" (in John Evans' terms[25]), thanks to Dick. Another important factor was that we were able to have a great deal of interaction with both Institute and LEAA decision-makers, to feel very clear in our minds about what kinds of information they needed to get from us and to be able to make changes in our plans in time to be responsive to those needs.

Our major frustrations came from impediments we had to face in the development of relevant information: restrictions on our travel and our presence in the cities, inability to collect our own data (i.e., reliance on the cities to provide us with data), and finally, problems of design arising from not having been involved early on in developing an evaluation plan for the program.

There are thus two areas where we'd still like to see LEAA move in terms of our findings and recommendations. The first one is a much more generalized application of evaluation planning at the national level; this still does not take place routinely at LEAA.

The second is the development of a more effective data policy. We feel, after the Impact program experience, that these two efforts taken together--better evaluation planning and better data--could significantly increase the payoff to evaluation in the criminal justice area.

---

[25] See page 111 above.

V. DISCUSSION (SPEAKERS AND PARTICIPANTS)

MR. GRANDY:

Shall we take some questions?

PARTICIPANT:

I'm Tom White, the Urban Institute. What happened to the other $20 million?

MS. CHELIMSKY:

That went for planning and evaluation. You mean the discrepancy between $160 and $140 million? The $140 million represents exactly what was spent on the action programs. The other $20 million were spread across the cities in increments of about $500,000 for planning and evaluation by the crime analysis teams.

PARTICIPANT:

Do you believe that those programs had an effect, or do you think it's just the luck of the draw?

MS. CHELIMSKY:

Are you asking whether I think they had an effect on crime?

PARTICIPANT:

Or any of those output measures where you believe there is a positive result?

MS. CHELIMSKY:

In those areas where we could get a really close enough look at the phenomenon, where we knew the process in detail and can explain

what we found in terms of the process, we feel we can talk about inferences, not effects. You have to remember, our evaluation was essentially a set of case studies. Where we were able to follow crime analysis team operations closely, for example, we do know what happened and we think we know why. Obviously we can't tell you that if we established that team again somewhere else, with different people and a different set of agencies, that the same outputs would be seen. But we do have the sense that we know pretty well "what happened," what the criminal justice problems were, and we have the evidence that the same techniques worked (or failed to work) in several places for reasons that we could document through close attention to the process.

PARTICIPANT:

I am Charlotte Moore with the Congressional Research Service. I just wondered whether you thought the criminal justice evaluation state-of-the-art is at the point now where it can be depended upon for making Congressional policy decisions? You may or may not know that your study was used by the House Subcommittee in its consideration of high-impact funds for cities.

MS. CHELIMSKY:

In comparison to what? Evaluation can certainly make as good a contribution to policy as other types of analysis presently in use. I think there is no doubt about that. But in terms of definitive inputs, in terms of "truth," I have to join some of yesterday's speakers who made the point that evaluation was just one part--a rational part but still only a part--of decision-making. Rather, we are developing evidence which should some day cumulate in better knowledge.

One thing I would like to reiterate about knowledge in the area of crime, and which doesn't appear to be adequately understood,

there are very serious data problems in terms of being able to say something about what has happened in terms of crime, whether, in fact, crime rates have risen or declined. We can't look across cities, that is, compare crime rates from one city to another city, because of differences in police tactics (involving more or less enforcement of the laws, for example) and because of differences in police and in victim reporting. What we have now is different people measuring rates of crime and recidivism in different ways, so we can't really say what they are or compare them across jurisdictions. In some cases you can't get the data. In others, it may be inaccurate. From the viewpoint of Congress, this is a major problem for judging the effectiveness of anti-crime programs.

PARTICIPANT:

I am Daniel Wilner from UCLA. Eleanor Chelimsky, I wonder if you have given thought to the generic problem of data? Yesterday I think we heard from someone from the National Institute of Mental Health[26] that there was reliance on the information gathered by the local community mental health centers. You are saying now and bemoaning the fact that, I think, there is a lot of variation in how information is gathered across the cities in the crime and recidivism field. I guess we can multiply the same problem for every area of inquiry in the evaluation field. Have you given thought to the generic issue then of local data and how it's to be used and demanded?

MS. CHELIMSKY:

I have given thought to it, but I don't know the answer.

I think what you can do in local areas is to require a lot more rigor in the record-keeping that people do, develop a lot more

[26]See pages 144 and 145 above.

understanding of how the data are going to be used (for administrative, management or evaluative purposes). I think all of those capabilities can be vastly improved from what they are now. But there are always going to be problems looking across cities, looking across projects. Even if you have similar projects, you are going to have tremendous variations in the way administrators administer projects. All of those things are going to mean that what is true in one place may not be true in another, and that we really need to know what the data signify in each instance before we can put instances together and examine a strategy--even with much better data than we now have.

I think the aggregation of data is the major problem we face because of local variation. I think it's extremely hard to say that aggregated data means something.

MR. WILNER:

Isn't there a need for some kind of national data policy or strategy in this?

MS. CHELIMSKY:

I think there is. We need to work on that.

PARTICIPANT:

John Greacen from the Police Foundation. I'd like to make an observation and ask a question. In terms of our discussion yesterday about the usefulness of evaluation, it seems to me this evaluation experience is very much in line with the kind of conclusions that I drew from yesterday's discussion. The Impact program as such was terminated by agency action long before the results or even preliminary

results of the national-level evaluation were available at all.[27] At the time, I thought that was sad. It seemed to me that evaluation should shed some light on that decision. I now see that that sadness was not necessary at all. Of course those decisions have to be made, and the challenge is to use the evaluation and its result in additional planning, which LEAA has been doing.

The question has to do with another issue. That is one that I find very troubling in the LEAA program, and I thought it was unique to LEAA; and now after yesterday, I find there are other agencies that have the same problem. LEAA is given a mission to enhance the capability of local and state agencies as well as to do things at the Federal level. The Impact program was specifically intended to do that, to create a planning and evaluation capacity at the local level and thereby to improve the performance of local criminal justice agencies. There is some very complicated mix of what can be done best through national evaluations or evaluations at a Federal level and what can best be done through improving the capacity of state and local agencies to do their own kind of work.

What lessons do you get from the Impact program on that question?

MS. CHELIMSKY:

It seems to me that the research gap I was talking about earlier is really what dictates the answer to that question because the issue

---

[27] Editor's Note: There may be some misunderstanding here since the Impact program was only slated to endure over two fiscal (or three calendar) years and did in fact last throughout its expected duration period and longer. The "termination" action to which the participant refers can only have been the announcement by LEAA in January of 1974 that the program would, in fact, be extended through June of 1975 as regards the crime analysis teams, while Impact projects and programs could continue to be funded until September of 1976.

of who does what evaluation, as between local and national efforts, is presently driven more by level of expertise than it is by the appropriateness of the organizational or governmental locus. To improve the interaction between national and state and local evaluations requires you first to improve capabilities at the local level. But how much effort is needed and what will be the payoff to that effort? Why do you need to improve their capabilities, in other words? In the criminal justice area, there is an assumption that improved research or analytical capabilities will result in reduced crime. We know we can't prove that this is so, presently, but most of us believe it. Impact cast little light, I think, on who should do what research but it did show that local capabilities could be improved. That, I guess, is why we were interested in the results of the COPIE-cycle--that it could be done, that it was feasible, that the cities did it and got a lot out of doing it. It's a policy decision whether the ability to do local evaluation is worth the cost of improving local research capabilities. I think it is, but the important question is whether you can get a result that is meaningful to you, not in procedural terms, but in relation to the substantive outcome you are trying to achieve.

VI. THE EMERGENCY SCHOOL AID PROGRAM (ESAP II): AN EXPERIMENTAL DESIGN

ROBERT L. CRAIN, Senior Social Scientist,
The RAND Corporation; and

ROBERT L. YORK, Program Analyst,
Office of Education, Department of Health, Education and Welfare

MR. GRANDY:

I think we should move rapidly on to our next paper which will be presented by Robert Crain of the RAND Corporation and Robert York from HEW. Their paper concerns the Emergency School Aid Program. Bob Crain has been at the RAND Corporation since 1973. He has his doctoral training in Sociology, but prior to that, he was trained in mathematics and engineering. Before he went to RAND, he taught at Johns Hopkins and he did this evaluation while at the National Opinion Research Center in Chicago.

Bob York is a project coordinator at HEW, he was formerly the project coordinator for the Coleman Report. He has done quite a bit of work in evaluation and planning in the area of school desegregation activities within the U. S. Office of Education.

I think that the first speaker in this team will be Bob Crain, and he will then turn the microphone over to Bob York.

MR. CRAIN:

Bob and I are going to talk about the 1971-72 evaluation of the Emergency School Assistance Act, the program of Federal funding to provide assistance to desegregating schools. The program was then called ESAP, with a "P", not ESAA, because the legislation had not been passed. In 1971, the program was keyed almost entirely to the

South because that is where all the desegregation was. It was a program which provided a fairly small amount of funds--averaging out to about $10,000 for every school that participated--which could be used to do almost anything that the local people thought was the right thing to do to help school desegregation along. I can be fairly brief in describing the project, in part because there is a paper in The School Review, entitled "Evaluation of a Successful Program: Experimental Designs and Academic Biases," which is on the table outside and available. That will tell you a fair amount about the program and the evaluation.

Just briefly, this evaluation is unusual because it has a genuine experimental design. The districts applied for funds with proposals to the Office of Education. Those that were funded, if they fell into the evaluation sample, were told, at the same time that they received their funds, more or less, "Congratulations on getting the funds, but don't spend them until we tell you to." The district superintendent was then asked to list the schools that he wanted to receive the ESAP funds in pairs, pairing them however he wanted to in terms of similarity. Those pairs were then randomized (coin-flipped); 100 elementary schools and 50 high schools were designated control schools, and the superintendents were told, "You may not use these funds in those schools." This happened in a hundred different school districts across the South.

It's a very simple, "after-only" randomization design. In the fall, there were randomized pairs of schools, with funds awarded to the treatment half of each pair and not to the control half. In the spring, the National Opinion Research Center came in and administered questionnaires and tests. Differences between the two groups, treatment and control, could be attributed to the program because of the randomization.

I should add as a footnote to the earlier conversation between Eleanor Chelimsky and Dan Wilner that the National Opinion Research Center collected their own data in all cases here. Every school district out there administered achievement tests. The Office of Education has found at considerable cost and pain that it's much safer to just start over and retest the kids than it is to try to use the local data even though in many cases the local data would be quite a bit better (a longer test and so forth).

Let me talk about the high school side of the study which is where the interesting results came out. When we came in in the Spring, the treatment schools and control schools were different. The treatment schools had more human relations programs going on. They had more in-service programs for teachers. They had more curriculum changes being made that year. The teachers in those schools said that the school was less tense. They said there was more discussion of race relations. The Black students in the schools said that their teachers were more sympathetic to integration. They were less likely to agree to the statement, "I feel like I don't belong in this school;" and they were more likely to agree with statements like "I like school."

Finally (and for many people, most important), the achievement test scores for Black male 10th graders in the treatment schools were somewhere between three-tenths to maybe five-tenths of a year higher in the Spring than the control group. Those are the kinds of results that are quite clear, and it's my feeling that you simply don't get that clarity without randomization. Mr. Seeman said yesterday that you can't take the nice, beautiful techniques we have in the laboratory out into the real world. But look, that is exactly what we did. The Office of Evaluation actually told a hundred and fifty principals and a hundred superintendents in a hundred school districts, "We're sorry.

The experimental design comes first. You get the money for this school but not that one." And they pulled it off.

You couldn't do that with some programs. I think the question of when you can do it and when you can't is an extremely important discussion which somebody should start.

I want to point out one other thing, which is that the result is a result only for Black male students. As far as I know, this is the first time a major evaluation had split the data by sex. If you stop to think about it, combining males and females is probably never a good idea, since they react in a social situation at that age very differently. Their whole relationship to school is quite different. But if the sex split hadn't occurred, the finding in the experimental design would not have been statistically significant. It wouldn't have appeared. We would have lost it. So that is important.

Another plus for the study is that the questionnaire was good on the race relations side, much better than preceding studies had been, I think. Perhaps part of the reason for that is that Bob York is the best person in the Federal Government on school desegregation research. He is in John Evans' shop. One of the advantages of Evans' shop is that it creates a situation where you can develop highly specialized professionals. And Bob works fairly steadily on school desegregation and has for quite a while. It paid off in this case.

I came out, at the end of the project, a fervent believer in randomization. But it has its problems. It is true that what randomization does is tell you that the treatment did indeed have this effect because there is no other explanation except sampling error. However, the treatment is nothing but money. Obviously, handing $10,000 to any school in the United States at any time will not cause

a rather sharp increase in achievement test scores of Black male students. We had to then start picking it apart, and figuring out what it was that they really did with the money. What were the local conditions that caused it to pay off? And there are some details to the puzzle which don't work out very well. Basically, the idea that seemed to come out of the experimental design is that ESAP created a situation where there were more human relations activities, more teacher in-service, more curriculum change, more concern about race relations in the school; and this spilled over probably into changing the motivation of Black male students, causing test scores to rise. Unfortunately, I derived a series of corollaries of the logical argument, and a fair number of them don't work. I don't know whether I have gotten noise in the data or whether the theoretical situation is so complicated that I didn't understand it. I think the latter.

Some of the serious problems with the evaluation are my fault. First, there wasn't enough emphasis on trying to figure out what ESAP actually did with the money. The paper that I referred to earlier argues that the reason why there was not enough attention paid to analyzing what happened to the ESAP funds is because the principal investigator in the study was absolutely and unequivocally committed to the proposition that there wasn't a chance in the world this program could work: and he wasn't going to waste precious resources chasing this damn thing around. That is what the paper says.

We have been talking about objectivity. But as it came up yesterday, objectivity had to do with an agency protecting itself. We researchers are the good guys, the agency the problem. But there are other kinds of objectives and there are other kinds of biases. In this case, the bias I brought to the project was a lot more dangerous.

I subscribed blindly to the shared ideology of the intellectual left, that authority is evil and institutions incompetent. I "knew" this program wouldn't work because everything the government does is wrong. I also think I wanted a null finding in order to prove to the world my independence, my "objectivity." And if it hadn't been for the experimental design, I probably would have succeeded.

At the end of the project Bob and I did a "dog and pony" show in which we said two things. First, this program is effective in terms of high school Black male students' achievement test scores. That is clear.

Secondly, we think it has to do with the emphasis upon human relations in this program, but that is not as hard a fact. We think it is true, and we have an argument that we can piece together. We believe it enough to tell it to you, but we don't have the kind of evidence we'd like to have behind it. At the moment we said this the program was in the process of being shifted rather drastically away from race relations and human relations toward remedial programs. What in fact was going on is that we were in the middle of a very big ideological brawl between the cognitive people and the social people in educational planning. The cognitive people felt that the need out there arose from the fact that Black students did badly on achievement tests; therefore somebody should get them to do something about it, and if you could indeed do something about that, everything else would fall in place. These people were opposed by other people who believed that the social relationships of kids--with each other and with their teachers--was somehow terribly important. We had done the kind of evaluation which people concerned with social relations would do in the sense that we had tried to measure the quality of

human relations in the school. And we were able to say in our presentation that it looked like the human relations thing made sense. But that begins a long story which Bob will tell.

ROBERT L. YORK:

Bob is being much too self-deprecating. He deserves a lot of credit, and in fact all the credit for a fine set of instruments in that study. One of the issues which John Evans talked about yesterday is, how do you implement the results of an evaluation study, and John mentioned the Policy Implications Memorandum which is a procedure for making specific recommendations involving action steps to be taken by various people within the agency.

With the Policy Implications Memorandum, I will talk about one recommendation which follows from the results that Bob Crain discussed. The Commissioner of Education agreed to a recommendation to increase the emphasis on human relations activities to some proportion (such as 30 percent) of the total funds. The recommendation was agreed to by all parties. The program office in fact had already taken one step by the time the memorandum finally got around to being signed. They distributed a memorandum to the regional offices which were responsible for the administration of this program explaining these results and explaining that they wanted more attention focused on human relations programs.

After the memorandum was signed, they also incorporated in their regional training programs the information that the Commissioner had agreed to this increase in human relations training. All that was well and good, but unfortunately, as far as I have been able to tell from

the evidence that I have seen, this process was not effective in changing the compensatory education and remedial orientation of the program.[28]

Why was that the case? I did not monitor or attempt to monitor the program office. They had been clearly in favor of the recommendation. They had not been in favor of this thrust towards compensatory education and the prospects for some success therefore seemed to be reasonably good. The recommendation could have been monitored by tabulating the amount of each ESAP award which was allocated for human relations activities. In the aggregate, 30 percent of the funds should have been allocated for human relations activities. This would work only in theory. If you put pressure on someone to reach a goal and they provide the figures to measure whether the goal is reached, you can be sure that the final figures will show that the goal was reached.

One factor which ran counter to our recommendation was the high percentage of repeat grants to school districts. This program had been in place for at least a couple of years, and many school districts already had established emergency school aid projects. The difficulty of changing project direction at the local level, after you have even this much of an established program, is pretty radical; and no doubt we underestimated it.

The recommendation also ran up against (although it was not totally inconsistent with) former Secretary Richardson's decision on compensatory education and back-to-basics which Bob Crain talked about.

---

[28] Editor's Note: That is, the orientation of the "cognitive people" referred to earlier by Robert Crain (see page 238 above).

The Policy Implications Memorandum process, at least the way I used it in this particular case, was too "top down," although there were meetings with the program office. Similarly, the program office itself took a top-down type of approach in its distribution of memos and centralized training sessions for the regional offices.

Finally, it is probable that the changes in program regulations needed to reflect a wider discussion and consensus in order to actually accomplish something. Parenthetically, the Act is tied in considerable--in fact gory--detail to regulations. The prospects of accomplishing changes in these regulations in a reasonable period of time were not good. The Office of Education, Head of Legislation and our lawyer, who must be relied on when you come to changing regulations, were not overwhelmed by this kind of evidence. The lawyer had gone on record previously as opposing any priority ranking of activities as being contrary to the detailed specifications of the law. So when you start trying to change policy, it clearly gets very messy.

A larger problem may be the limited nature of policy recommendations that are likely to follow from overall impact evaluations. The thing that an effectiveness evaluation does best is to tell you whether the program should or should not be funded. This study, although much more encouraging than most, was still ambiguous in answering this basic question. Impact evaluations also analyze program components associated with a favorable outcome. The human-relations program effect was one such example.

While other, more ambiguous, program effects were found, there were none, other than the human relations effect, to recommend to policy-makers.

Where does this lead us in our subject of uses of evaluation? As some speakers suggested yesterday, and this morning,[29] I suggest it leads us to participate in planning activities with program managers. This exercise hopefully helps the program by clarifying program objectives and also provides the evaluator with a basis for developing an appropriate evaluation. When this planning effort seems to be reasonably successful and a new or revised program seems to have a fairly well articulated set of objectives, an effectiveness evaluation may well be a good evaluation strategy. Where there is less reason for optimism, however, an effectiveness evaluation is not likely to be of much use. Ambiguous results about the overall effectiveness and program component effectiveness are highly likely and will not address the real problems which lie in the legislation and/or the administration of the program. If a program is lacking in clear objectives, even with the able assistance of an evaluator, there is pretty high probability that it has not articulated a model or a mission. At worst, it will be all things to all people, a program that has built a constituency but lost an identity.

Under these conditions an evaluator may provide the best guidance to the program by an evaluation that provides a few elements. Before discussing these elements, let me point out that an evaluator's participation in planning activities may make his objectivity questionable, creating a potential conflict of interest situation in view of program staff, particularly if he has fought a few battles and lost them. In such a case, I would suggest the evaluator use this valuable experience to write the work statement for the Request-for-Proposal, or whatever procedure is used in specifying the design of the evaluation, and then turn the evaluation over to a colleague. I would not simply have the evaluator pull out of the picture because

[29] See, for example, pages 112 through 115, and pages 214 and 224-226 above.

I think one of the crucial mistakes that we make in a lot of our evaluations is not getting in quickly enough at the beginning; and the planning activities that an evaluator may participate in may be very helpful in designing a sensible evaluation right from the start.

Let me conclude now by listing a few of the key elements in a completed evaluation of a program that seems to lack direction. First, the program's manager must be convinced that it is true that the program lacks more than fancy objectives stated in management-by-objectives language. Evidence must be shown, if it is true, that there is confusion and lack of direction in the program. This leads the evaluation to the tedious task of reviewing proposals that are submitted from, in this case, local school districts from all over the country. It leads to interviewing Federal program staff at all levels. If the planner-evaluator is correct, this process will show how confusion in the direction of the program has had an impact on the technical assistance offered to applicants and on the ambiguities faced by those who review those proposals and decide who gets awards.

Second, there should be site visits to the grantees. These will probably document the lack of direction of the grantees, and some method should also be provided--and there are lots of ways of doing it--of assessing impact, although the method used would almost certainly be much cruder than the elaborate methods (such as the ones in the study that Bob just talked about) typically employed by effectiveness evaluations.

And third, the evaluation must provide some specific substantive guidance for program managers. The program staff that was unable to provide substantive guidance before the evaluation will be unable to do so if the evaluation only documents what is wrong. There are doubtless many strategies. I will mention two that I have used.

One is to rely on the existence of several successes in the projects that are site-visited and provide enough detail in the report on these successes to give guidance to the program on what makes a success. This limited case study type of evidence is crucial in my judgment. Put another way, evidence based on statistical analysis of desirable project characteristics is not understood or trusted by program managers. Short case studies which contain essential elements of success give program managers much more information and more evidence that th contractor's understanding is deeper and does not reflect what they view as simple statistical manipulations.

Second, if you doubt that there are enough natural successes in the program, the evaluator may design a study with what will euphemistically be called comparison groups. These comparison groups are projects which are not necessarily Federally funded, and which will be selected in some way to increase the probability of success for site visits. The case study type of evidence presented to program managers under this option is essentially the same as that I mentioned before.

In conclusion, this type of evaluation strategy, agency interviews, site visits to grantees and a design that provides substantive guidance for success, offers a good prospect for agonizing reappraisal and constructive direction in such a reappraisal. I think that a combination of factors can help make this more than a paper exercise. The program managers I deal with are, in my judgment, people of good will who have genuine commitment toward the goals of the program in which they are working. If we learn to work with them more effectively, I think that we will have more successes than failures. Thank you.

MR. GRANDY:

Thank you, Bob and Bob. Let's take a few minutes here for some questions.

VII. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

PARTICIPANT:

My name is Gordon Bermant. I am with the Federal Judicial Center. We are very concerned in the judiciary now with the concept of experimentation with regard to court processes because we feel that there are enormous legal and ethical problems that arise when cases are assigned at random to treatments. This is the first time I believe in the meeting so far that explicit mention of randomization was made. It struck me that one of the reasons it worked was because of the relative powerlessness of the people receiving the money. You could put that on them without their fighting back.

There are many kinds of evaluations you'd like to do where you just can't do that, where people just won't stand still, if they know a randomization is going on, for being the control group. Perhaps they are justified in exercising whatever power they have to thwart the value of randomization.

Do you have any general comments on the relation between scientific integrity and power relationships in dealing with this kind of issue?

MR. CRAIN:

Yes, I have thought about it. It is certainly not true that school superintendents are powerless in dealing with the Office of Education. They normally walk all over OE. This is a situation in which OE moved way out on a limb, scared out of its mind, and pulled it off. It reflects a big commitment on OE's part to take

a big risk. A lot of power is in commitment. The fact that OE was committed this time and had never been committed before--that, in itself, changed the power balance.

There are very important ethical questions about randomization. For example, in this particular case, Bob York and John Evans could go to the program people and say, "Look, it's a very small program. We are giving you the total amount of money that you would have received anyway because that is all that's appropriated. If we randomize, certain kids by the flip of a coin don't receive it. But if we don't randomize it, a large number of kids aren't going to receive it anyway. Unless you can argue that the kids who got randomized out are somehow obviously more deserving than the millions of kids who are not being served by this program anyway, why is it such a big deal?" And that argument, I think, eventually carried some weight. It was one that struck me as being quite ethical.

If you have a program which is serving everyone, then you have to argue that the treatment and the control group are both receiving something which reasonable men would say is equally likely to be useful. In this case, you can't just give the control group nothing. The control group gets something that you believe may not be as effective as the new idea; but you can, with good conscience, say that there is no evidence that my new idea is better than the old idea, and therefore, there is no evidence of real discrimination. Indeed, if we don't implement the new idea, everybody is going to get the old idea, so everybody is going to be discriminated against.

There is a paper by Donald Campbell, "Methods For the Experimenting Society," which goes deeply into this. There are some conditions where it clearly cannot be done; you clearly could not randomize Title I. Title I is a very large program designed to

reach every impoverished child with a fairly large amount of money. Depriving children of Title I because of the needs of an experiment would seem to me to be unethical. But I think there are lots of cases where it can be done. And I think there are lots of cases in the criminal justice system.

PARTICIPANT:

Ben Liptzin from the National Institute of Mental Health. Ont of the things that we have learned in health evaluation, particularly in terms of drug trials, is the necessity for <u>two</u> control groups, one receiving a placebo. You mentioned the fact that the experimental design showed that the treatment was effective. But isn't it possible that it was something analogous to a placebo effect? I wonder, for example, in your design, whether the control schools were also notified that they were going to be part of an evaluation in terms of consciousness-raising organization of the community interest in the program? In order to be able to separate out what was effective--money itself, versus identifying the school and triggering some changes, don't we also need to know, given what happened with the money, that a superintendent didn't try to do some of those things in other schools in the district, even the control schools, to screw up your design, if it seemed like a useful thing to do and didn't require too much money?

MR. YORK:

I think on the question of what effects there may have been in the control schools, we don't really know, of course. But we did <u>not</u> notify the control schools. We attempted to make as little a deal about that as possible. There was some data collection, of course, in the school, but to the degree that an issue was made of the fact that there were control schools--that was something we did not impose.

There was an oddity about the program which worked, I am convinced personally, greatly to our benefit. That is the funds got there very late, and I think that prevented the superintendent from getting his act together and moving some Title I funds around so as to compensate for it. I also think that we worried an awful lot about that happening. I think we were a little paranoid. In fact, every school, a typical school in the inner city, receives 20 different Federally-funded grants. Nobody can ever sort that out to make it equitable. I don't think most superintendents try terribly hard. They are making a conscious effort, but they are not going to kill themselves to see that every school gets exactly the same nickel.

The business of placebos is tricky when you are dealing with human relations within a social organization because it is very hard to distinguish logically between what is a placebo and what is motivation, which is what you are trying to produce.

MR. BLOCH:[30]

Peter Bloch from the American Bar Association. I just was thumbing quickly through the report while Bob York was talking, and I'd like to ask Bob Crain a question about the methodology. I noticed in quickly looking through the report that you used regression analysis to find your results, and that suggests to me that you thought perhaps there were background differences in the experimental and control schools. Could you comment briefly on the reason you used regression analysis?

[30] Member of the Research Perspectives Panel.

MR. CRAIN:

There are some background differences which persist despite randomization. They are not statistically significant and would therefore be normal in a randomization. We took these out by a multivariate analysis of variance.

There is a great deal of attention paid to regression analysis in the report, but that reflects the point I made earlier--that since I was absolutely convinced that the experimental design wasn't going to work because the treatment could not possibly work, I was not going to waste any time on that. I was going to try to do something interesting so we wouldn't be throwing the Federal Government's money away. So I ran multiple regression equations by the ton, all of which produced nothing except gibberish, more or less.

PARTICIPANT:

My name is Evie Rezmovic. I am from Northwestern University. My question relates to the level of treatment imposition necessary to obtain a desired result from an evaluation. It seems that the ESAP Program was a vast effort--I think $64 million was spent on the program. You said that there were $10,000 spent on each high school. Apparently there were 300 schools altogether, grade schools and high schools, included in the study. Now I am not sure how many students were included, how many were attending each high school; but if, say, there were a thousand in each high school, it might break down that the amount of money spent per student was $10.

The results that you got are fine, of course. What I'm questioning is whether you could maybe have gotten more or greater results had there been some kind of greater treatment imposed, had there been

more money spent per student. How does one determine how much treatment is needed? There are a lot of problems that come up having to do with whether evaluators ask the right questions. How do you define the problem? It seems that a related important issue was how much treatment do you actually give to get whatever outcome you are looking for?

MR. YORK:

Those are good points you make. The problem is that when you start an evaluation, you start collecting cost data. It gets extremely complicated. I agree that is an important policy question to get into questions of whether there are linear effects or not by costs. But when you do that, when you make that decision, you are clearly adding a great deal of money and a great deal of effort to the data collection.

Secondly, a lot of these kinds of programs that we are talking about, of human relations types of activities, do not tend to involve large sums of money. So $10,000 in a school in one sense, if they are not buying huge numbers of remedial curriculum materials and so forth, but focusing on rather straightforward training is not necessarily a small sum of money.

MR. GRANDY:

Thank you. I think we will go on to our next presentation; and after that, we will take a short coffee break.

VIII. MANAGING INVESTIGATIONS IN ROCHESTER: AN IN-DEPTH CASE STUDY

PETER B. BLOCH, Staff Director for the Commission on Law and the Economy, American Bar Association

MR. GRANDY:

The next paper is going to be presented by Peter Bloch. He is an attorney and is presently affiliated with the American Bar Association. His paper, however, concerns some work he previously did while at the Urban Institute where he worked from 1968 to 1976. This is a study of the police investigation system in Rochester.

MR. BLOCH:

I'd like to start by explaining that my situation is a little different from that of most of the other people here because I have left the field in which I did the work I am going to report on. I'd also like to explain in advance that I will say some things that are going to be critical of the Law Enforcement Assistance Administration, and I am going to do so with some apology to the people who are present, because, unlike some prior commentators who dislike bureaucrats, it seems to me that most of the bureaucrats I have known have tried their best, and that the problems often are problems of management and leadership, more than problems of bureaucrats who are lazy and resistant to change and who can't accomplish things.

I am going to try to set one evaluation of the Rochester system of managing police investigations in the context of the Federal Law Enforcement Assistance Program even though it was done for the Police Foundation not for the Law Enforcement Assistance Administration. I'd like to start out by commenting on something that has been said

many times before today, but not in the same words. That is that evaluation is a support system. It works in support of management. If there is no management, there is nothing to support. If the program knows where it's going, if it has some ideas of what it is trying to accomplish, then it may be possible to work with evaluators to get information which is needed by management and can be used by management. That requires, of course, that there be some communication between people with management skills and people with evaluation skills so that reasonable requests for information can be made; and information will not be requested or provided if it is not likely to be used by management.

Often the Congress is blamed for creating conditions which make effective evaluation impossible. It is said that the goals or programs are too vague or inconsistent, and that therefore, the programs can't be run adequately, we can't have clear objectives, and we can't do evaluation. That seems to me to be an interesting criticism, but I prefer our Constitutional system of Government to others. I think there are problems in a Congress. It is a collegial body. The goals for agencies are never going to be very clear. There has to be an interaction between the Congress and the administrators of programs. The administrators have to get their acts straight also and to take the responsibility for devising reasonable programs within the statutory framework, using a combination of management skills and political skills--because you have got to keep your fences mended with the Congress.

The most key management skill that I can think of is one suggested by Richard Neustadt in his analysis of the Presidency, in which he suggested that before a President undertakes a program, the program managers should figure out how they are going to get from here to there. They should figure out how they are going to implement

the program. If they can't do that, if they haven't figured out how they are going to accomplish the result, they might consider whether or not they would like to accomplish it. They should think twice about doing an evaluation of a program if they do not know how it can achieve its expected result.

Generally, the LEAA program presents some of the problems of other programs for the Federal Government. But to some extent, it is among the most inconsistent of programs. On the one hand, it has the goal of giving block grants to states. On the other hand, it has the goal of requiring the states to follow in detail a planning process which was set up by the Federal Government. These are somewhat competing and conflicting aims, to my mind. It makes it difficult for the Federal Government to implement an effective program. It seems to me that thought should be given to the extent to which we really do want to give money to the states, and then give it; and thought should be given to the extent to which the Federal Government should exercise a leadership role, and in those areas the Federal Government should accept that role. But to be continually fighting with the states to follow paperwork requirements and to engage in confrontations over plans when there are no serious Federal objectives seems to me somewhat doubtful for an effective program.

In the area that I did my research, which is police investigation, LEAA has funded several pieces of research and has contributed something to the knowledge of criminal investigations. The first important piece of research was done by Bernard Greenberg at Stanford Research Institute; and in that research, he documented a fairly simple but important fact that if the managers of police investigations examine the reports of the preliminary investigation conducted by patrol officers, they can determine the likelihood of success in

individual investigations if possible investigative resources are invested. Police can be somewhat more effective if they stop investigating cases where there is a low likelihood of success and continue investigating cases where there is a high likelihood of success.

Another piece of LEAA-sponsored research was by the RAND Corporation. I am going to simplify a little bit what the RAND Corporation report found, but I am going to also give you my own interpretation. The RAND Corporation was a study of the state of the world. It was conducted primarily with questionnaire, used to find out the structure of police organizations along some predetermined dimensions and to determine some effectiveness measures the police departments could supply from data available to them--despite the fact these data, of course, are known to be dirty. It was found that when you examined the relationship between the structural dimensions that RAND had identified in advance and the fairly dirty measurement instruments, that there was no detectable relationship between methods of police organization and the effectiveness of the investigation effort of an individual police department. That does not mean that you cannot manage a police department so as to be more effective in criminal investigations. It only means that RAND was unable to detect the ways in which that is or may be done.

I also did some work for LEAA on managing criminal investigations. Don Weidman and I completed a study which was published as a prescriptive package. Ours used a case-study technique. We went to six police departments, and we found essentially what RAND found, except that we described in detail what each of the departments was trying to do, so that there were some suggestions from individual departments, based on their experience, of logical, rational management ways of trying to improve police criminal investigations.

The study about which I intend to talk most today is the study of managing investigations in the Rochester system. What happened in that case was that Tom Hastings, who was the Director of Planning of the Police Department in Rochester, came to the Police Foundation saying that he had an innovation which seemed to improve the quality of investigations in the Rochester Police Department. He called the innovation, "coordinated team policing." It consisted of assigning some detectives to work together with patrol officers in a single unit at the street level, commanded by a police lieutenant. This is different from most police departments, which take great pains to separate their patrol division (usually found on the main floor of the main building) from the detective division (which may typically be found on the third floor some distance away, sometimes with its own luncheon facilities so that the patrol and detective officers need not talk frequently to one another).

The idea behind coordinated team policing was that it would be helpful if the people who started police investigations would talk with the people who were going to continue those investigations. They could get to know one another, trust somewhat the quality of one another's work, perhaps avoid the unnecessary duplication which occurs when the police detective goes back and asks the citizens exactly the same things that the patrol officer had asked--either because he never got the report from the patrol officer in the first place, or because he has the attitude that all patrol officers are dumb people in the first place and that there is no use in ever accepting the value of any work from them.

What happened when Tom Hastings approached the Police Foundation is that he presented clearance statistics which showed that somewhere over 40 percent of Rochester's burglaries and an unusually high proportion of robberies were being cleared. The statistics were so favorable that they were greeted with some skepticism at the Police Foundation, which thought, perhaps with some justification, that statistics of that sort only came out if there was something funny going on in the statistical system. Now, the Police Foundation is an interesting organization because it is run by an ex-police commissioner, Patrick Murphy, and has a board of directors whose members are very active in policing. It also has a staff which is working regularly with police departments. So it has some knowledge of what police people think are important operational questions in policing. It identified the report from Tom Hastings as an important report worth further investigation, but it specified a two-stage process in order to conserve the research resources which would go into it.

Frankly, I was extremely skeptical of those statistics; and I expected that the first phase, which was an audit of the books in the Rochester Police Department, would discover that the results were due to the way the statistics were kept, and that they were not due to actual operational differences in the police department.

Our first report, called "Auditing Clearance Rates," examined several ways in which those statistics might have been jimmied. For example, we compared the arrest records, before and after, of the officers who were in the teams--both the patrol and detective officers, because the results might have been produced just by assigning better quality personnel to the experimental treatment. We examined reclassification practices because it is possible that the police were more ready to determine that things were not crimes which existed in

the experimental area, thereby reducing the denominator and keeping the numerator (i.e., the number of cases cleared) the same, thereby increasing the clearance rate in the team area.

We also examined the multiple clearance question (i.e., how many cases are cleared for each case for which a person is arrested) because the criteria for determining how many cases to clear are somewhat subjective. In Rochester, they were particularly subjective because Rochester used a rule of clearing cases based on a judgment as to whether the suspect had committed offenses other than the one for which he was arrested; and that judgment was reached by using the personal judgment of the detective who had made the arrest in the first place. There was little supervision which would have reduced the number of clearances claimed as a result of an arrest.

Basically, having examined those and some other possible sources of error, we determined that in Rochester there was no bias either in favor of the teams or against them. Therefore, further investigation was warranted.

In our follow-up report, called "Managing Investigations, the Rochester System," James Bell of my staff, who is co-author of this paper, lived in Rochester for over a year, which is not exactly hardship. But it did enable him to know the people in the police department and to get some understanding of whether there were hidden factors which perhaps would not be disclosed to someone who just walked in from the outside and did a three to five-day study to find out whether an exemplary project was in existence. He was there, and he lived with the police department.

We then did manual checks on the records, coding original reports from the records to find out the quality of the investigations which were conducted and to track the reports through to see how many investigations resulted in arrests. As a result of that tracking, we found that the Rochester system seemed to produce more arrests for robbery and burglary; and we believed that we could attribute that improvement to the program. We also had one finding which troubled us somewhat and suggested management controls were needed, and that was that there was a somewhat smaller success in court with on-scene arrests in the team areas than in the non-team areas, suggesting a possibility that the teams had become somewhat more aggressive in their criteria for making on-scene arrests. (Although we were aware as well that the team areas presented demographic characteristics which might have made it more difficult for the police to maintain witness cooperation and to obtain success in court.)

The most promising feature of the Rochester system, I believe, is that the detectives were placed in the teams under the control of team commanders who then managed the case investigation process using, in part, a system like the one that SRI had documented in California. The Rochester system had been developed independently, within the Rochester Police Department, to close cases which were not promising, using the detective-lieutenant to assign cases or investigative tasks to individual officers in order to capitalize on the special expertise of individual team members.

After these studies were done, LEAA held two conferences. One was a conference with evaluators, and another was a conference with some police chiefs. The conference with evaluators resulted in a number of suggestions for how a demonstration program might be designed to find out more about criminal investigations. The

conference with the police chiefs was not designed to help construct a program to find out more about criminal investigations. It was primarily for informational purposes to tell the police chiefs what LEAA had found. In fact, there is a national demonstration program in team policing which attempts to follow-up on all of the pieces of research which I have discussed here. However, it doesn't do that very well.

One problem with the demonstration program is that the RAND Corporation believed that, as a result of its study, reductions in the number of detective personnel would have very little effect on (i.e., would not hurt) investigative success. I think their basis for believing that may have been somewhat flimsy, but it might well have been a possible ground for further investigation. It was not included as part of the program. Resource differences in investigation are not being examined by LEAA.

Our study suggests, I thought, that it would be helpful to do a demonstration program where detectives and patrol personnel worked together closely in patrol units, since we found that that had a promise for being a successful program. That also is not part of the demonstration program. The demonstration program consists primarily of a training program which is trying to get police officers in local departments to conduct better preliminary investigations and which is trying to attend to some of the system problems of the criminal investigation system. I think it's an interesting hypothesis. Of course one of the problems is that it will be hard to duplicate the training program that is now being constructed. Furthermore, there was no advance indication that a special training program would be particularly effective in this field.

One thing that troubles me about this follow-up by LEAA is that, in my mind, the improvement of the police investigation system is essential to the improvement in local policing. It dates back to the case of Mapp v. Ohio,[31] in which the Supreme Court decided that police officers had to get information in legally, constitutionally permissible ways; and there was a hope expressed by the Justices of the Supreme Court that police departments would find ways to get information in constitutionally permissible ways.

In light of the patrol experiment done by the Police Foundation, and also in light of close analysis of the likelihood that aggressive or preventive patrol by police officers will produce improvement, I think that the single most constructive approach to improving the contribution of police to the criminal justice system is by working on ways to improve the collection of information from individual citizens, the apprehension of criminals and the prosecution of criminals in court; and that ought to be a major emphasis of the LEAA program. Enough resources ought to be devoted to test alternative hypotheses. To test them, LEAA should find police departments willing to implement programs that promise success. Then, LEAA should work with police officials and with local prosecutors to design a program which will implement the program which was chosen for experimentation. You don't easily graft things onto police and prosecutors. They should be part of the design process.

There should be a commitment in advance that the programs participating should implement specific experimental programs. That, in fact, is not the case in the present demonstration program, resulting

---

[31] 367 U.S. 643, 81 S.CT. 1684, 6 L. Ed. 2d 1081 (1961).

in still another case study analysis which will only give us further hunches about what hypotheses we should then test to find out what works.

In their design of the evaluation of this program, the organization chosen as the evaluator makes this quite clear. The evaluators are going to study, first, whether the demonstration agencies receive and interpret the technology being transferred under the auspices of the Managing Criminal Investigations Program, how the sites plan to integrate the technology into ongoing operations, what components of the technology were actually implemented in each demonstration site, what was the impact of the implemented technology during the demonstration evaluation period, and whether impact can, in fact, be attributed to the program. Given the fact that a similar program has been drawn for neighborhood team policing, apparently without successfully implementing the program as orginally designed, there is little reason to believe that the full Managing Criminal Investigations Program will be implemented at each of the sites. We therefore are likely to find, in this much smaller program than the one Eleanor Chelimsky talked about, that there also will be different programs at each of the sites, and that the evaluation will consist primarily of case study judgments about what happened.

I think in this area we need a commitment to finding out what works in the managing of criminal investigations, and we haven't started doing it yet.

Briefly, I would suggest that LEAA, in designing programs, ought to work more closely with the people who are going to implement those programs so that the operational people will accept the programs

when they try to implement them. That is part of the leadership process in which local governments can be drawn into implementing programs which may work.

Where there is no leadership plan, it seems to me that we might be better off to seriously consider backing off by not requiring a mixed, internally contradicting process of planning and block grants. Instead we should give money to the states or to localities with the most serious crime problems. Then local governments will be accountable to their own people for the way in which money is spent.

The last thing I'd like to say is that one of the most important problems in running the LEAA program (and many other programs) is the problem of time. Unfortunately, our political officials tend to have fairly short time horizons, and good programs take long periods of time to implement effectively. The need for time requires statesmanship on the part of our public officials, because it is much easier to design a program which may help even a little bit in the long run. It also takes confidence for an administrator to believe, when he is designing a program, that even after he has left, there will be other people willing also to act in a statesmanlike manner and to continue worthwhile programs once they are started.

MR. GRANDY:

Thank you, Peter.

IX. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

MR. GRANDY:

We will take a few questions if you have some for Peter before our break.

PARTICIPANT:

I am Judd Kenney, Department of Justice. Actually this one perhaps spans both of the presentations, those of Ms. Chelimsky and Mr. Bloch. Recently, the Attorney General has proposed a separate organizational entity which would be exclusively devoted to the compilation and reporting of crime statistics. My own liking would be a Census Bureau for Crime Statistics.

Now, from Ms. Chelimsky's efforts, one could derive an affirmative attitude toward such an organization. Now, addressing Mr. Bloch's Rochester study and its outcome as far as LEAA is concerned, would you view LEAA as having a continuing role as an evaluator of programs and the new organization as we understand it--let's say, superficially--as merely having an accumulative role and a reporting role; or could you two get together some idea of how these two efforts would interrelate? or would LEAA be out of the program of crime data and evaluation?

MR. BLOCH:

The single most important role that I see for LEAA is in research, demonstration and experimental evaluation. I think that is a very important role for it to continue to play in an improved fashion. The data collection agency idea starts getting at an important

problem, but I don't think it gets at it well enough. This is something I feel strongly about. The fact is that after over eight years of planning in 50 states, we still don't have good documentation of the flow of offenders, except perhaps in one or two states.

It seems to me that the public interest requires that when we are talking about agencies that deal with liberty and safety and equality, that there is a very strong interest in public information about the individual actors in that system. So I would prefer that there be requirements that the disposition records before individual judges, the disposition records by individual police units and by prosecutors, the recidivism records for types of offenders and for different races and backgrounds of offenders--that this information be collected and be a matter of public record so that we can not only identify where the problems in the system lie, but we can also try to hold our criminal justice officials accountable for their contribution or lack of contribution to the success of the system.

PARTICIPANT:

My name is James Bell from the Urban Institute. I have just one question for Peter. Where do you see compelling proof in the research that has been conducted in criminal investigations that it is important to move detectives, in other words, to create organizational trauma to patrol in order to achieve improved investigations? As I know it, we have one piece of research that suggests that. We have no other empirical proof. For us to sit and decide that programs should be designed to include that element without that kind of proof is, I think, premature. I guess I'd like to know what substantiates your basic dilemma with the now-constituted Managing Criminal Investigations Program?

MR. BLOCH:

First, I must point out that Mr. Bell was my co-author on this study. He is the man who spent the time in Rochester.

I'd like to say first that it's my impression from the results of that one study which was in only one city, that there is a good chance that the detectives working in the same unit had an effect. I also think that on policy analysis grounds, on thinking about the way that criminal justice systems work and the way police departments work, that I am convinced there is good reason to experiment with that hypothesis.

I would emphasize that I didn't say that my hypothesis should be selected by LEAA. I only suggested that LEAA should work together with officials in the field to develop programs. I believe that if they do that, that they will find there are a substantial number of agencies which, when presented with the evidence and when persuaded to take part in a program where there is leadership at the Federal level, will be interested in participating in a program in which it will be possible to find out whether assigning detectives to teams will have an important effect. I personally believe that it would have an effect.

MR. GRANDY:

Any other questions or comments on this topic? Okay, we will take a short break at this time and then resume in about 10, 15 minutes.

X. THE NATIONAL EVALUATION PROGRAM: KNOWLEDGE SYNTHESIS

JOE N. NAY, Senior Research Associate,
The Urban Institute

MR. GRANDY:

Joe Nay is going to present his perspective on knowledge synthesis. Joe is currently at the Urban Institute. He is an engineer by training, a graduate of a joint program between the Electrical Engineering Department and the Sloane School of Management at MIT. He has done quite a bit of work primarily with interdisciplinary teams to alter the operations and improve the effectiveness of large organizations, both inside and outside of Government. His experience covers management problems, policy research, practical problems of implementation and also evaluation. Joe, it's a pleasure to welcome you.

MR. NAY:

After listening to everyone else yesterday, I reworked my talk last night. I don't know if I have done a good or a bad job yet; but I'd like to start with something that happened to a friend of mine a few years back, which, I think, puts some of the things you heard yesterday in perspective.

This person decided to do a series of interviews with high-level analysts and high-level policy people in a series of departments in the Federal Government. He collected a lot of names from many of us. He interviewed a lot of the analysts and I was very interested in how it came out because I had been close to the past work of several of those analysts. A lot of their work had had effects that I knew about.

Some of the effects were very positive. Some of the effects (I thought) destroyed things that I was very fond of. I had a lot of mixed feelings both about how the effects of their work had come out and how all of this would come out in the interviews.

I think that both of us were astounded when he came back with the first round of interviews. Almost universally, people in these analysis and staff groups had told him that they hadn't had any effect at all. I plowed through some of the interviews myself with him. I even found that some people whose effects I knew of (because I had been working with line management people at the time the effects of their work took place) had said, "The most frustrating thing about my three-year tour was that I didn't have any effect at all." How could they say that?

I sometimes think that people in staff groups and a lot of evaluators and analysts, in particular, have a vision in their head that is left over from "Executive Suite."[32] That serial has done more harm to management than anything else that ever happened. It left people with visions of <u>the big meeting</u> where decisions are made. Everybody has a cigar, and they say, "What shall we do?" The analyst reads off his numbers, and they say, "That's it. That's it. That is what we are going to do!" Few analysts ever actually find themselves in such a meeting; perhaps that is why they think that their work has no effect. If you look upon evaluation as gathering information to have an effect on an organization or upon the decision-makers in that organization, however, I think that you have to look very carefully at the sort of ripple effects that each effort has.

[32]Editor's Note: "Executive Suite" was a movie, genre soap opera, serialized on television during the fall of 1976.

In one sense, I think a lot of those analysts were right. They had often gone to meetings and taken their papers with them. They said, "This is what we ought to do!" And the decision-maker didn't do exactly what they said. But in some particular cases that I knew about where my friend found the interviewees still saying, "None of my stuff had any effect," I knew that in many cases it had had widespread effect, either by altering some course of action, or preventing another one, or by really sealing a choice that people hadn't quite made up their minds to make.

So I think that even the idea of effect is more in line with what Donald Elisburg said last night.[33] Whether something has effect or not depends upon what different people will accept as proof and how their actions are influenced, or bounded, by information that they believe.

The National Evaluation Program at LEAA is partly a knowledge synthesis program. It's broken into a Phase I study which is a synthesis and assessment study and larger Phase II evaluation studies. I'll talk a little bit about how that came about.

A Phase I study is really a synthesis of the information that is available. We could talk for hours about what I think is necessary and unnecessary to do knowledge synthesis, but I want all the Phase I grantees to stay in the room so I'm not going to give that talk. This way, the Phase I grantees won't have heard this entire talk already.

The important thing about the NEP (after hearing yesterday's high-level people from agencies around town) is that it is something that has been carried out. A lot of information has been gathered together. A lot of knowledge files have been built. It is kind of interesting to see how that worked. Our role is as technical advisor,

[33] See page 201 above.

and we are doing a case study of how it all happened over the last two or three years and how we think it all came out. We are also giving intermediate advisories along the way of things we think ought to be changed.

The present emphasis on oversight is one of the factors that is leading to the development of these syntheses programs in several agencies right now. And acceptance of the results hinges in part on degrees of proof. English is a funny language. There are two definitions of "oversight." The first one is supervision, superintendency, inspection, charge, care, management and control. A lot of people forget that there is also a second definition of oversight that is used every day, which is the fact of passing over without seeing, omission or failure to see or notice, inadvertence.

I want to talk today about a real life attempt by an agency to convert what a lot of people thought was a case of the latter definition to a case of the former definition, the National Evaluation Program.

When I used to try to teach people about evaluation in Government programs, I always required that they look at a program and find out some very simple things at the start. I used to keep pounding, "Go out and look and see if it exists." People say, "Evaluators haven't done anything." But there are hundreds of programs around the country that never were implemented in anything near the shape in which they were envisioned. And without evaluators, no one would ever have known this in many cases. I think the evaluators have pointed that out, and I think that is a valuable function. So the first question about a program is, Does it exist? and the second question is, What is it? What process is in operation? What is it that exists? What outcomes are produced (you have heard all these before in any evaluation paper that you have read) and what impact do they have?

We can't do any less for the NEP. There will be a case study out in May where we will try to answer those questions for the first two-and-a-half years of the program. But we can answer the question now (sort of from the laboratory to you) although we may have to reverse ourselves later. We can say, Does the NEP exist? Yes. What process and operation? We can't tell you all about it today in a half-an-hour, but we have it pretty well documented. What outcomes have been produced so far? Nineteen studies have been produced, and there are eight more underway. There will be another batch next year. What impacts do they have? Some of those impacts are being captured through surveys and interviews. Others won't be.

For a number of years, as a couple of people have remarked, the bulk of LEAA money went into the block grant program. The block grant program was originally, by design, a case of the second type of oversight. At one time it was characterized as "leaving the money on a stump and letting someone come and get it," the way people used to buy moonshine. This was a result of an argument about whether local initiatives or national categorical programs were better; and for a long time, LEAA had this block grant program. There were tens of thousands of grants out there, hundreds of most any kind that you could name that were commonly known. They were locally determined, and most of their evaluation, if it was done at all, was done locally. Most of the national evaluation effort was made against the discretionary money, on that part of the money that national LEAA controlled.

The 1973 Act required oversight in evaluation. If you can picture what happened, you go along for a number of years. You give away your money. People make grants with it for things that they think are good. Suddenly Congress says, "You don't know what they are doing. You don't know how it's working out. We want some oversight information about this."

Most people suggested that four or five big evaluations be done immediately, that large, long-term evaluations with clear assumptions be put in the field. The problem was that when all the internal suggestions were produced of what should be evaluated, there were (on the last list that I could find when I was preparing this talk) 122 topic areas that people had suggested as needing one of these five costly evaluations.

Many groups in Government have been faced with similar problems, and I think many groups have called in the universities and selected five topics and begun large-scale evaluations. Some of these have efforts worked out; but, as you heard yesterday, an awful lot of them have run aground. They have come back with findings about the nature of what is out there. What was being done in the field has been different than everybody thought. The measurements selected in advance by the agency and the evaluation grantee haven't exactly fitted the programs to be measured. There has been controversy about the results.

LEAA did, we thought, a clever thing. They convened a task force whose director is in this room and settled upon a strategy of trying to milk knowledge in sequential steps from those locally-determined block grants in order to go at it in stages and try to build some information files. A little over two years ago, they came to us and said, "We want to try one of your approaches of buying knowledge in sequential stages." That is always a pretty good thing. It makes you feel good if they say they want to try one of your approaches. The bad part was they wanted us to help. After a lot of hassling over the ground rules, we agreed to serve as technical advisors and to do a case study of what happened.

In the face of all of the same pressures and problems that were outlined to you so gloomily yesterday, of pressures from up above,

pressures to hide results, vagueness of objectives, certainly a lack of consistency in many of the programs, enormous gaps between theory and practice, the National Evaluation Program has come into being. It has produced the 19 Phase I studies that are complete and has 8 more underway. Despite the problems that you heard about from executives from half of the Federal Government yesterday, the full studies are available. You can get them. You can check them out of the library or you can get them on Microfiche. Some are better than others. You can get them all. Summaries of all are being distributed.

The summaries which are written by the grantees are nationally distributed. Some demonstrable impacts have already occurred, and we are following up with surveys and interviews to try to check out some more. Every study has been preliminarily rated, both whether it's the kind of thing we thought we were buying with Phase I work descriptions, and on what we think the apparent usefulness of it is. The program has been kept stable long enough that we are beginning to have a good idea of what some of its strengths and weaknesses are. Changes are now being made to improve some of the problems that have cropped up.

In May, as I said, the case study will be available; and you will be able to see what we think about the whole process.

In light of what you heard yesterday from various officials who told you why something like this cannot be done, it's hard to understand how this could have happened. So I've revised my talk on knowledge synthesis to try to outline for you here today the key things that I think allowed it to happen. I have five here. (There may be a different five in the report.) They are:

- Simplistic thinking
- Stubbornness
- A detailed approach
- Pressure to follow it
- A single person in charge

Let's see, simplistic thinking and stubbornness. I think people sort of outlined some simple things to do and they stuck with them for a year or two, an underlying concept or two that didn't get modified until the agency could begin to see how they worked. Unusual, but it happened. Two more key factors were the work description (i.e., a detailed approach) and pressure to follow it. I think the fact that a single person was responsible for it (Dick Barnes[34] who is back there in the corner and ought to be up here speaking) is major. He has stuck with this thing for two-and-a-half years. He has been responsible for it, and he has been the focal point for it. He has gotten encouragement and occasional discouragement from the heads of his agency and other people in his agency. He is still on the program. I think his strong determination to do these obvious things--read the proposals, look at the concept papers, talk to the grantees, try to get people to modify their approach a little bit so they come out a little better--has been a key factor.

One of the simplistic ideas was that too little was known about what was actually happening in many topic areas to really begin full-scale evaluation. This led to the idea of a Phase I, Phase II exploration. I will not talk about Phase II today.

Phase I is really a form of evaluability assessment, and we will talk a lot about the nature of what we think evaluability assessment is.

---

[34] Editor's Note: Head of the National Evaluation Program at the National Institute of Law Enforcement and Criminal Justice.

Phase II is a larger, longer evaluation where one appears warranted, and after you know enough about the area to better begin to scope one.

People talked a lot yesterday about the dangers in the evaluator's job. There are a lot of dangers in the evaluator's job, and I believe Jim Stockdill noted that the evaluators may often be the only persons who are looking at both the rhetorical charters and the operating activities.[35] From the standpoint of an organization trying to implement programs, you don't want to ever sell that activity short because questions about performance come from those rhetorical charters in many cases. The measurements that will have to be taken if an evaluator does the measurements himself will always be out where the activities are. When we talk about evaluability assessments, we are trying to assess that gap and bring the rhetoric and the activities closer together before buying major evaluations.

Again, you have heard my stories before. There is a favorite quote of mine in one of Shakespeare's plays that goes something like this. One fellow says, "I can call dragons from the misty deep." And the other replies, "So can I and so can any man; but the question is, when you call them, will they come?"

Now, various private and public groups have been busy calling those dragons from the deep in the form of policies and even programs to solve problems. It has only been a few years, really, since the Office of Economic Opportunity would end poverty, police chiefs would end crime, school superintendents would end reading and math problems, especially among the poor. The evaluator in many Governmental operations has been (for a number of years) the only person who was required to go out and see if these dragons came.

---

[35] See page 129 above.

By an evaluability assessment, we mean a design approach which looks at the project or process that is described by the people in charge, and looks also at the process that exists in reality. Trying to bring these two sectors together is an attempt to match up this measurable information with the questions, the goals, the objectives of the people in charge. It is true that you may find their objectives (not the people, of course) very fuzzy. You may find both the objectives and the activities very fuzzy. But by working with those people in charge and with the theory about what is supposed to work and how it is supposed to happen until the rhetorical purposes of a particular Government activity are reduced to a series of evaluable statements, you have half your problem solved. In many cases, we see evaluations where people then go to the field; and they try to assess (but there is a lot of argument in our own group about whether you should go to the field and assess at that point) whether those evaluable statements are true. If the activity in the field, on the other hand, is really quite different from the rhetoric, there are a lot of cheaper ways--than formal evaluation--of finding out how different rhetoric and activity are. A smaller, cheaper study where you try to collect that information is one of those ways. It is also a lot less visible than going out and doing a massive evaluation and finding out that the implementation is quite different, even though it may be either good or bad.

So the other half of evaluability assessment consists of recording carefully the service process or direct intervention that is actually being made and attempting to create a measurement model of the real activity of a project. This is carried out so that what is actually being done can be described in the most mundane and concrete way you can find. From this, you can assess what in reality can be measured, what those measurements would be, how they would be taken, how much they would cost and exactly where they would be obtained. By now, you anticipate my next step.

The end result of an evaluability analysis is an attempt to marry these two sets of information together and see if you can match up the potential answers that you can get with the potential questions that everybody is interested in.

We now refer to two new types of error. We not only have Type I and Type II errors;[36] we now also have Type III and Type IV errors as well.

Type III error is going out and measuring something that doesn't exist and coming back with numbers about it.

Type IV error is going out and measuring something very well, but not getting any of the things that anyone is interested in.[37] When you go to that big decision meeting in the sky or you try to distribute the information, you find that you have measured a lot of information about a real activity; but none of the things are interesting to the people who are in the discussions about what is to be done with them.

We will say if you only have two hours to design an evaluation, spend one hour on the rhetorical program and one on the actual direct

---

[36] Editor's Note: Type I error: the rejection of a true null hypothesis (that is, obtaining a statistic indicating there has been an effect, when there is no effect).

Type II error: acceptance of a false null hypothesis (that is, obtaining a statistic indicating there has been no effect, when there is one).

[37] Editor's Note: These problems are discussed at length in the Urban Institute's Working Paper 783-34, "Evaluability Assessment: Avoiding Types III and IV Errors," John W. Scanlon, Pamela Horst, Joe N. Nay, Richard E. Schmidt, and John D. Waller, January 1977.

intervention. If you have two days to design an evaluation, try spending one day on each job. If you have two months, spend one month on each job.

It is not so much that there is a fixed cost to evaluability assessment, but that there must be a fixed attitude of these continuously recurring attempts to match the answers to the questions and the questions to the answers. Because you are really trying to design a workable path for producing information out of what is going on and bringing it back to the people who are in charge of it. We put great stock, as you can tell, on bringing information back to the people who are in charge of it, even if they don't want it.

At the same time, you are really getting a lot of the basis for a technical evaluation design. We don't view this effort as a prelude to evaluation. We really view it as a use of evaluation tools in producing information, although people will make a lot of arguments about the level of belief; but I think those are philosophical arguments. There are many ways of producing things that are just beyond question (or beyond belief!). Unfortunately, a lot of those academically sure ways do not work very well in actual complex programs. There are a lot of ways of producing less convincing proof that can be applied pretty well. You are always in a trade-off between what is possible and what is desired in a real program and a real program evaluation.

The typical local criminal justice administrator needs to know more about a new approach than that outstanding people under a particular set of conditions (which are generally different from their own) were able to do it successfully. We believe that before gambling on an approach, an administrator needs to know if it has been successful in a variety of settings when operated by ordinary people. In this sense, the broad block grant program is pretty good. If you can collect a lot of these projects in a topic area and they're being operated

by ordinary people in operational agencies at the local level something may be learned, whether it's in the court or police or corrections or diversion programs. What did we send Phase I grantees out to do? The work description is available also.[38] Call Dick Barnes and get the work description. Somebody described it last night as a spiral staircase.

The NEP Phase I study tries to introduce a short intense prior step, a form of evaluation design that includes the synthesis of measurement models for the area under consideration, collection and assessment of the information that is available so you can try to see what is known, what will need to be known and what is knowable. Don't forget that last step. You may find yourself in a position of promising people answers that simply aren't knowable from the programs that exist.

By going step by step and exploring what is known, we feel that a quicker overview can be provided. Unnecessary errors can be avoided in design or evaluation, and a file can be created on a topic area as you go along. One of the toughest underlying concepts to implement in these studies grew out of evaluability assessment. A conscious attempt was made to meld together the theoretical thinking in a topic area, what actually occurs in field operations, and the methodologies of measurement and evaluation. Tom White, who is here today, says that most of the one-person problems have been solved. There have been enough bright people around long enough that most of the problems that one person can solve have been taken care of. A lot of the problems today are team problems. You don't find very many people who are awfully good in theory in a topic area and who are also good in

[38] Editor's Note: The Work Description for a Phase I Study is available from the National Institute of Law Enforcement and Criminal Justice, LEAA.

the measurement and evaluation that needs to be done later. You really need to meld those skills together.

We and the grantees--probably they more than us--have found it a very painful meld. We tried to do it with a structured work description that included issue papers in the area to try to address the theory and what people said was being done and should be done. Flow and function information from actual projects in the field was also included. First, a survey of the projects (usually by telephone) and then visits to a lot of projects to try to take down exactly what intervention was carried out and how it's connected to the criminal justice system. Then we ask study teams to synthesize a framework for description and evaluation and to assemble against this framework what knowledge is already available that has been produced in other reports and what knowledge they picked up on their field visits. In other words, they are to call out in terms of the framework and the issues what everyone wants answered, what gaps there are in the knowledge and how they might be filled. They are also asked to try to design the measures and the approaches they would use, if they had to look at a single project in this particular topic area. I will give you a list of topic areas later, but they are quite diverse. The work description had to be fairly general.

There was a lot of argument at the beginning about how much this should cost and how long it should take. Arguments ranged from $20,000 in four months to hundreds of thousands of dollars in years. We finally settled on a kind of a nominal size which varied little with the different topic areas. LEAA shot for a six- or eight-month turn-around which proved to be, I think, too optimistic; and certainly most of the grantees who are here will feel that that was too optimistic.

We kept track of it all as they went along. After running the first batch through and looking at them, we knew a lot more about the

process. Each of the 19 full reports completed has been read by all of the members of a team made up of LEAA and Urban Institute people, and each has been rated as to its Phase I-ness and probable usefulness. We have kept at it until we have gotten forced-choice paired-ratings on several criteria. As Dick said in one of our meetings, "It's a very select game. In order to come to the table and play, you have to read all 19 reports." One of the reports is 1,800 pages long. Some of them are shorter than that.

The early Phase I study leaders' comments and complaints were all gathered and combined. We took a lot through interviews and a lot through meetings that we had at different times with people doing the work. These were combined with the ratings of the study, section by section, to try to get information to rework the work description. When one of these things goes right, you are not exactly sure what has happened; and when one of these things goes wrong, you don't know quite whether you made an error in explaining it, whether the topic area is sort of impossible, or whether the grantee has fallen on his face. With a sample of 19, obviously I'm not going to say we have experimented and will determine the critical five or six factors that are in there. But I will say that we are keeping track of them, and we are trying to feed them back now into what the agency is doing so that they can do a better job on the next ones that they do.

We are using phone surveys to follow up the summaries that are distributed. I didn't bring any summaries with me, but there are small summaries that are being distributed nationally. We are doing phone surveys of local and state people to see, did they get it? Did they read it? What did they think of it and can they tell us anything they have done as a result of it or anything they are going to do?

CONTINUED

3 OF 4

We are using interviews to follow up actual users in the agency. There are several of the studies that have actual line users in the agency, and we are going to interview them. We already have done some interviews to follow up what they think they got out of the study. So we are trying to put all of this together and address this question of joint levels of use, the question of what is effective information to put out. There is one thing that people were saying yesterday which is very true--that the higher you go in an agency, the more people want and need one-line descriptions. When Congress improves their oversight, this problem will, of course, go away. They will be ready to take complicated textured information about textured programs. But until that happens, the higher up you go, the more you need something that is almost a press-release level of information about the study. I think it has been very hard for the grantees because they know that their information is going to be reviewed at various levels and they can almost predict at different levels who is going to be happy with it and who is going to be unhappy. Nevertheless they have gone ahead drawing up their summaries. And LEAA took a policy quite early that not only did they not want to affect (if they could help it) what their grantees put in the summary as far as conclusions were concerned, but that they didn't even want to give the appearance of affecting it.

The Urban Institute reviews each product as well as LEAA. If we think the summary doesn't match the content of the full report we send them an advisory, and we say, "Hey, we don't like this part of the summary because we don't think it matches what's in the report." There is a regular process for convening, meeting and having an argument about that. But the further up you go, you do have to reduce the amount of information; and there is more and more pressure to have a result that matches what people previously told people they

are doing and what people previously told people the results are. However, the grantee's own final summary is made available in each case.

We have some difficulties in deciding how well we are doing in terms of study quality. If you let 19 studies and you know what you'd like to get out of them, how many of them should be good? We do have informal knowledge of other people's internal reviews of sets of studies where somebody in some agency has looked at the research that they have bought. Generally, if 50 studies are examined, say, some of them can be eliminated. That is, they are not any good at all. Another batch of them may have usefulness, and another batch of them are really useful. Generally, the figures that I have from various agencies run about 35 percent, if you want to take a middle range of how many studies turned out. That is, 35 percent of all studies let are really useful. Unfortunately, not enough of these studies of buying research have been done systematically, and not enough have been done in an open way where you can use them for comparison. There is still enormous pressure on people in Government to say that every grant that they let produces something.

If you are not in Government, you can say, I am going out and I am going to let 50 grants and I expect two-thirds of them to go sour. If you are in industry, you can do that with your research; nobody expects all your research to pan out. But in Government there is still this feeling that all grants should be perfect; they should all come out. If anyone here should happen to know of any comparisons that I can use on yields of contract research, I wish you would see

me some time in the next few days because I only have one or two comparisons now that I can publicly use. Three or four that I thought I could use have been withdrawn by people who called up and said, "Gee, when I gave you that letter, I gave it to you for your own use; and I really don't want you to use it as an open comparison because nobody here will understand." We are really having trouble grappling with that issue of what kind of yield you should get out of a set of studies like this. We are going to try to treat it in the case study, so if you all have examples that you know of, that I can use for comparisons, I'd appreciate them.

I will just run through the topic areas of the first 19 Phase I studies. They were Neighborhood Team Policing, Specialized Patrol, Traditional Patrol, Crime Analysis, Pre-trial Screening, Pre-trial Release, Youth Service Bureaus, Prevention of Juvenile Delinquency, Juvenile Diversion, Alternatives to Juvenile Incarceration, Detention of Juveniles and Alternatives to Its Use, Project IDENT, Citizen Patrol, Citizen Reporting, Early Warning Robbery Reduction, Premise Security Surveys, Treatment Alternatives to Street Crime (which is a drug treatment referral program), Court Information Systems, and Halfway Houses.

Let me anticipate a question by saying that when we took this approach to the topic of Prevention of Juvenile Delinquency, nobody thought that anyone would come out with a complete framework for juvenile delinquency prevention. But it was an area of examination that was just getting on its feet. The agency had to have some tools to go in and explore it. Because this was a structured approach, they pushed some people into it to do some early exploration from which they could use the data and information that were produced in their continuing work.

I think that is about all. I am ready to open up for questions.

XI. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

PARTICIPANT:

Don Weidman from OMB. Joe, in talking about your factors of success, you left out Joe Nay. I am wondering in a program like this, we can sort of understand what Dick Barnes and his people do. We know what grantees do. But what has been your role and that of your people, and how important is that to success? And do you think the typical agency can understand and accept giving a contract to people such as you to somehow or other assist bureaucrats in doing what they are supposed to be doing anyway?

MR. NAY:

We are going to treat that in the case study. One of the questions to us that is important is, can an agency run something like this by itself? In other words, this is a development, really, of some of our ideas. Can we develop it to the point where it has more applications without our help? I don't know if I know the answer to that yet. We are accumulating a lot of information. If Dick were near a microphone, I'd make him answer. I think there are a couple of functions that we serve. We are in a position where we can go back to a meeting of agency people two months later where everybody at the agency is harassed and under pressure to do something else, and we can keep saying, the thing you decided you were going to do was this. Don't forget we have written it down, we are still writing it all down. Think how this is going to look in the case study. You said you were going to do one thing and then shifted it.

I think we help provide some stability for the program in that sense. I think we probably have provided more of a push for orderly process, for review and for content assessment than I see in a lot

of other agencies (remember some of the negative speakers yesterday). I think having somebody outside who is looking over your shoulder or who is arguing with you maybe pushes that a little harder. I don't know the answer yet, I think.

MR. BLOCH:

Joe, one model for trying to get good research is the one you suggested where you carefully specify what is going to go into the research, and you compare things against what you have specified. Could you comment briefly about the role also for what could be called duplication of research, but could also be called competition, in order to produce products in fields that are identified in advance as important?

MR. NAY:

I pushed at one time for multiple studies in each topic area. I am very much in favor of that. If you pick a topic area and you say this is really important (the National Science Foundation, as you know, did some multiple studies), you should put a couple of teams on each one, at least. There is some competition between them. There is a good chance that one of the studies will get done and get all the things that you want, or maybe one will get one part and one will get another. You can synthesize it inside the agency. I guess if I were in a vacuum and I had my choice, I probably would use dual grantees. It certainly would make my job a lot neater and easier because then I would always have at least one relative comparison. I could say this grantee did it. This other grantee didn't.

Dick or Jerry, do you want to comment on that? How comfortable would you feel having four grantees out studying the pre-trial release process?

MR. CAPLAN:

I am not familiar with the NSF experience. I think it would be very difficult for us, and I would be cautious about it.

MR. NAY:

I think you would take an awful lot of heat from many directions at present for funding multiple studies of the same topic area. You would have to have an awfully clear press release at some point which spelled out why this is a wonderful thing.

MR. GRANDY:

Thank you, Joe. I think you let him off awfully easy there with the questions.

XII. THE KANSAS CITY PREVENTIVE PATROL EXPERIMENT: A FIELD TEST

GEORGE L. KELLING, Director of Field Evaluation,
The Police Foundation,

and

JOSEPH H. LEWIS, Director of Evaluation,
The Police Foundation

MR. GRANDY:

Let us go on to the final paper from our researchers with a presentation by two gentlemen from the Police Foundation. Their paper concerns the Kansas City Prevencive Police Patrol Experiment. I think it's one of the better and more decisive experiments that has been done. The participants are George Kelling, who is Director of Field Evaluation at the Police Foundation, and Joe Lewis, who is overall Director of Evaluation at the Foundation.

Joe has, I know from personal experience with him that dates back to the early 1950s when he was in the Weapon System Evaluation Group in DOD, quite a lengthy background in evaluative research work.

They are going to speak in the order in which they are listed in the agenda, with George going first.

MR. KELLING:

For those of you who will notice how I visibly wilt during this presentation, I will only say that I have one of those colds that I have only experienced since my 40th birthday. There is something unique about these colds. It seems that every injury I ever experienced in my body comes to the surface again--the ankle I broke in

my 20's and the rib I broke in my 30's, both start aching again. Over the last three days, I think I have spent three-quarters of my time in sleep. The rest of the time I have been reading Humboldt's Gift which again lulls me to sleep in my reading chair. This is my first venture beyond my living room. I don't say that by way of begging sympathy if any of you want to attack anything that I say. I say it by way of apologizing for not being as unpleasant as I normally am as I make my presentation.

In 1971, Clarence Kelley, through a bond referendum in Kansas City, suddenly had 300 new patrol officers. At the same time, the Police Foundation had $30 million. The Police Foundation didn't know how to spend its $30 million, and Clarence Kelley said that he didn't know how to use those 300 new officers. So they invited a group of experts, all of whom were from the Police Foundation, to discuss what should be done with those 300 new patrol officers. At that point, I knew what an expert was. An expert was a person who other people thought was an expert and who didn't deny it. I didn't deny it. I sat there and everyone thought I was playing the role of the village idiot. It might have been that I was the village idiot because it simply didn't dawn on me what to do with those 300 new patrol officers. It turns out the vast majority of the people there, including the command staff, couldn't decide what to do with those 300 new patrol officers. Some of the police officers thought that it would be best to decrease the size of beats. Others thought you could have two-man cars. Others thought you could use them in highly technical ways. Any time we turned to the literature for guidance about what seemed to work and what didn't seem to work, there just was very very little evidence.

The result was that the command staff and the experts couldn't decide what to do. It was kicked down ultimately to the patrol officers.

The patrol officers were told to decide what the problems are in their areas and what to do with new officers. In one of the areas, the South Patrol Division, a vigorous debate began. They decided the most serious problem was dealing with youths in the area, and they wanted to divert resources from patrols to deal with that youth problem.

One half of the task force said, you can't do that. Preventive patrol is so important that you simply can't divert resources from those purposes. Another half of the task force took the stand that, no, preventive patrol isn't that important. "We are bored out there half the time. There aren't that many calls for service, and who knows that it all makes any difference anyway." Out of that disagreement grew what has come to be known as the Kansas City Preventive Patrol Experiment.

Preventive Patrol, the movement of police vehicles around some kind of geographical area, generally has two primary purposes. The first purpose is to create a feeling (and this is the view of O. W. Wilson, the former Reform Police Commissioner in Chicago) of police omnipresence. That is, by the movement of vehicles through a city, you could create a feeling, both on the part of citizens and on the part of the bad guys--that the police were always present, that they were always around.

The second purpose of preventive patrol was to put police vehicles in places where they could rapidly respond to calls for service. The idea behind the need for rapid response to calls for service was that that would lead to more apprehensions of criminals at the scene of crimes. And that would lead to increased citizen satisfaction. That would have all kinds of good and wonderful results. Indeed, most of that seemed to be logical. It was logical that if you put police vehicles in areas, had them move around swiftly, that that presence

would deter crime. It was logical that if you got response time to particularly low levels, that you could increase arrests and apprehensions as a result of that and that citizens would feel better about that.

With that in mind, we divided a 15-beat area with a population of 142,000 in 32 square miles into three experimental conditions. The population density was something like 4,000 citizens per square mile. It ran the gamut of everything from an all-Black area out to a white suburban area and the wealthiest area in Kansas City. It was an interesting cross-section of the population.

What we did was to match the beats in triplicates (that is, we had five matched triplicates of beats) on the basis of calls for service, crimes, other demographic variables and then randomly selected from each of the triplicates one beat which would be called a proactive beat (I'll define this later), another group which would be called the reactive beat and a third group which would be the control area. We called the experiment The Proactive-Reactive Preventive Patrol Experiment.

We called it that because we hoped that would fill the first paragraph in any newspaper article about the experiemnt, and nobody would be willing to read much beyond that. In other words, we had an issue on our hands: we were going to suspend a public service that is considered essential. We did not want the public to know about it, and we decided that a long time in advance. We worked to make sure publicity releases, etc., were far enough in advance that by the time the experiment was actually operating, citizens did not know about the experiment; and we did hire players (fairly high-class Black pimps) to go into Kansas City to find out what street criminals (Black street criminals at least) knew about the experiment going on there.

(It turned out that although they discovered a great deal about the operations of the Kansas City Police Department, they did not know about the experiment itself.)

What we did in five of the reactive beats was that any time a police car was not in response to a call for service, that police car was to leave that beat and go to an adjacent beat (or the closest beat) which was designated as proactive. Let me be very precise. We manipulated only one thing in the experiment. We manipulated the amount of time that police had available for preventive patrol in an area. That is all that we manipulated. We did not manipulate any other strategies. It was not aggressive patrol. It was patrol as usual, except that in the reactive area, the police left the area as soon as they were done with calls for service.

In the proactive areas, we increased the number of police and we added cars. We added other conspicuous police vehicles to increase the level of time available for preventive patrol to somewhere between three or four times the amount of preventive patrol. In the control areas, we left everything the same.

Given that the goals of preventive patrol are to reduce crime, to increase citizen satisfaction and reduce fear, to increase business persons' satisfaction and reduce their fear, as well as manage traffic, we measured all of those variables. We measured crime through reported crime, and we measured crime through a business persons' survey. We measured arrests in the area. We measured attitudes through community surveys and a business persons' survey. We had observers, four civilian observers for a full year; two police observers for a full year; two other observers for two months all to observe what was going on (to notice how the officers behaved, how much cheating was going on, etc.). We measured traffic. We also had the observers interview

citizens that had actual encounters with the police so that we would get measures of citizen attitudes from people that actually had had contact with the police rather than those that did not have contact with the police.

We started the experiment in July of 1972. That was after a little over a year of planning. We started it with great festivity, and we were really going to have an experiment. Two months later, it was very obvious to everyone concerned that there simply was not an experiment.

We had set certain conditions whereby police officers could go into their beats. For example, one thing they could do would be serve warrants. Warrants have never been served in the history of Kansas City as they were during that time. Warrants were being served over and over again. It was obvious that the officers simply were not living up to the experimental conditions.

Also, it turned out that there weren't enough cars available for the proactive area. We had excess officers riding in an extra car.

We decided to replan, and we very gently stroked the officers explaining what we had in mind, why it was important. Chief Kelley came and explained why it was important. We went out of our way to try and make sure people understood. The task force itself went out and explained. Also for those people that deliberately sabotaged--well, I won't go into all of that, except that there was one Lieutenant (and those of you who know policing will recognize this), there was one Lieutenant who was transferred to the 11:00 to 7:00 shift at the jail. For those of you aware of the status hierarchies of police departments, you will realize that something was being said to him.

We restarted the experiment in October 1, 1972; God, that was a long time ago. We conducted it for a year. It took us about a year-and-a-half to get the final report out. The summary version is available on the table outside. The total technical report is a thousand pages long, and in it we present all the details of the problems that we had with the experiment, along with a good share of the data for people who want to look at it themselves. We have also had requests for the data, and Northwestern University and several other places now are doing secondary analysis of the data that we have.

In total, we made 640 comparisons between the three areas. We found statistically significant differences 40 times. We used a .05 level of significance.

There was no consistent direction in the findings. That is, the proactive did not always have higher scores or the reactive lower scores or the reactive higher scores. There was no pattern in the statistical significance found.

In sum, as a result of the experiment, we concluded that the level of preventive patrol simply did not affect those variables that we measured in Kansas City at that time. The degree to which that can be generalized to other cities has to be determined by people in other cities, and we provide extensive demographic information about Kansas City in the report to other cities that would like to consider the implications of Kansas City for themselves.

I should say that, on the other aspect of preventive patrol, response time, a very major study is now being done by LEAA, conducted in Kansas City as well. (We have put out a relatively minor study in comparison to LEAA's.) Our study, which measured citizen satisfaction is very interesting, and one that I think has delightful policy implications. That is, it turns out that the most important determinant

of citizen satisfaction with police service is not the length of time it took police to get there. Instead, there is an intervening variable, the citizen's expectation of how long it takes the police officer to get there. In other words, if the police officer is expected in three minutes and gets there in five minutes, the citizen is dissatisfied. (That is an exaggeration.) But if the police officer is expected in ten minutes and he gets there in five minutes, then the citizen is satisfied. In other words, expectation intervenes.

LEAA's study which is coming out shortly presents a surprise finding that I think many of us simply didn't think about. That was, the length of time it takes citizens to call the police. The police go to the scenes of crimes like gangbusters. Citizens take long periods of time to call the police. That is not surprising when you think about it. It is not surprising at all.

After we were done with the study and we talked to patrol officers and we talked to patrol officers in many many cities, they told us. "We are not surprised by either of these findings. We are not surprised at all."

I'll turn it over to my colleague and leader, Joe Lewis; and now I will completely wilt.

MR. LEWIS:

I want you all to understand that I know we are all the way over the hump. One of the last conferences I attended, someone volunteered the information that in rating schoolteachers--students were doing the rating--they got some results which, at first, they couldn't understand because they didn't seem to have anything to do with the characteristics of the teacher as commonly measured or of the students either for that matter--subjects taught or anything else. It just turned out

that teachers who taught a class that occurred at 10:00 in the morning were always rated highly. We are way beyond that. I have been way beyond that for some years now.

I was asked to talk, following George's talk, about "What happened then?" after the Kansas City study. One measure of success of an experiment and evaluation would be that you look around a little while later, and everyone is doing it. You'd say that is a good thing.

First I should say that our point in doing this, from a national perspective, was that we thought police administrators would benefit from knowing that it's a good idea to question what you do. Secondly, that police agencies can do it and this Kansas City experiment was an excellent demonstration of how much police agencies can do if they are oriented in a certain way and given a little help. And thirdly, that there are resources available to them by which they can test more directed kinds of activity. If they think they know something that is likely to have more effect than routine preventive patrol, they should go ahead and try it. You can take the resources from that, and you won't have to worry very much about anything extraordinarily bad happening. That was what the experiment suggested. These were the things we hoped for.

But let me now talk about the population into which we wished to plant those ideas. It's highly fragmented, as I think you all know. The nature of knowledge about it is illustrated by the fact that three or four years ago, people said there were 40,000 police agencies. Then a survey was made by Census, and it turned out there were 25,000. We lost 15,000 in no time. Now people think there might be 17,000. I don't know where the other eight thousand went, but anyway 17,000 is still a very large number. Its importance is that policing in general is a very insular occupation. Police people are to a degree

insulated from the society that they are policing. I think that is all familiar to you. They feel very defensive about it. They are very secret about what they do. But from our viewpoint in trying to inculcate knowledge and get a cross-jurisdictional notion of what might be useful, a most important point is that they are also insular with respect to each other. People rise in the police hierarchy almost entirely in the first one that they join, or at least in the one that they stay in. They are only beginning to share knowledge with each other. Lateral movement of practitioners is very rare. It only occurs in most states at the level of chief and not very much then.

There is no long tradition, in policing, of research. There is no loop from academe to practice and back again as there is, say, for city managers. They are all, nowadays, college-trained in a rubric they all understand. There is a great deal of movement from city to city, upward in size or complexity. There is a growing profession which feeds on new knowledge and which expects that new knowledge is not always useless. Sometimes it might even be helpful. The police are very suspicious of research and very suspicious of people who do it. It is something that is done to them. One has to say at the outset that there is some justice on their side. It is very hard to point to much research that has helped them. I am sorry about that. I wish it were not so.

So in the face of that kind of population, what was it reasonable to expect? Well, one of the worst things that can happen to anyone doing studies is that nobody knows about it. We at the Police Foundation are very careful and put a lot of energy and planning into the orchestration of the release of information from our studies. The Kansas City study is a good one to think about from the point of view of saying, Did anything happen in consequence of it? because at least it is well known. A lot of things people know about it are not true, but nevertheless it is well known.

Unfortunately, however, it has another characteristic which complicates things very considerably. It is not a prescriptive study. It doesn't say, "Stop doing this and do this instead." Many people have complained to us, who rather like the study otherwise, that it doesn't tell them what to do. It doesn't in any specific way. Therefore, what do you think ought to happen?

We like to start at one end of that. There is a sort of continuum. The first thing you'd like to do is see that people are thinking about it. They not only know about it, but there is converse about it, people are paying attention to it, perhaps asking themselves questions about their own enterprise.

Well, we can say that in the case of the Kansas City report, there has been plenty of discussion. It continues. Often it is violent. That does not disturb us. We think that's very good. People are paying attention when they are arguing. And the applicability of the information contained in it or the modes of thought that it suggests will settle into place in an appropriate way in time, along with the additional work that is being done by LEAA, and other things which will add to that body of knowledge. That is the first thing to know. There has been a good deal that happened in those terms.

To go to the other end of that, what did Kansas City do itself? Kansas City spent about 18 months of planning, real research and study and program planning, to come up with a directed patrol, a set of strategies which they are now testing with LEAA funding. The reason I mentioned 18 months is that it shows it is extraordinarily difficult to say, "Well, if you don't do that, what do you do? How do you accumulate in usable pieces those fragments of time in which you normally would be doing preventive patrol?" It's a very difficult problem. It has been addressed in Kansas City. They were deeply devoted to it.

They believe the study is what I am saying, and they have paid attention to its indications. They do have a program to test other ways of using those same resources. We find that somewhat hopeful.

In New Haven and half a dozen towns around it which constitute a planning area in Connecticut, they state that what they are now doing, which they call directed preventive patrol, is a direct outcome of their consideration of the Kansas City study. I am not familiar enough with the program to know how to judge it. But I do know that their patrol activity being centrally directed (that is, the objective of it being shifted in accordance with crime analysis activity that they think tells them what the current problems are), is satisfactory to them. It is a direct result of this study.

In between, in San Diego where we have also done experimentation, but not on this subject, the whole department is being converted, or has been converted, to what they call community-oriented patrol in which they give the patrol officers a great deal of responsibility and freedom to take the time in which they are not responding to calls not in riding around, but in learning very deeply about their beats. They are expected to learn in a formal sense, and their learning is measured in terms of demography, economics, land use and so on, but also, in more subtle ways. They are asked to concentrate not only on what the problems are that may lead to order maintenance difficulties or criminal activity, but also what are the resources in their territory that could be used to help them? Is the fellow who runs the drug store on the corner good with the kids? Is that a way of dealing with the problem? They say that they do that in preference to preventive patrol because they think it ultimately will make a more effective, more concerned, more relevant set of police activities.

There probably are others, but those are the two or three cases of concrete action that I happen to know of. In spite of what I said

about the state of affairs when one thinks about "Let us change policing," some things are going on. There is a good deal of discussion, controversy and thought and some cases of actions to test other ways to use time more usually spent in routine preventive patrol.

It will be an awful lot easier, however, when some of the trends that are now showing up--increased education, particularly management training for police managers--have had more time to take effect. It's beginning to grow a little bit. Use of special analysts and people of that sort who don't grow in police agencies and have to be brought in from outside, which opens up the police agencies to wider possibilities for improvement--when those things continue, it will become easier; but we have a very long way to go.

XIII. DISCUSSION (SPEAKERS AND SYMPOSIUM PARTICIPANTS)

MR. GRANDY:

Shall we take some questions?

PARTICIPANT:

I am David Smith from HEW. If you spent your time doing this experiment, which is quite an interesting experiment, and Kansas City is still spending 18 months, or whatever it is, trying to figure out what to do with these 300 patrolmen, why did anybody give them to them in the first place?

MR. LEWIS:

I think you are being swept away by logic. It's well to guard against that. Or let me put it differently. One needs to get into the frame of reference within which that decision was made. That decision was made right after a riot was quelled. Does that help?

But the question that they were addressing on directed patrol was not what to do with some additional officers, but what to do with a segment of patrol officers' time, which represents in the aggregate some $2 billion worth of national resource. It's a thing that might be worth working on 18 months.

PARTICIPANT:

Tom White, the Urban Institute. Do you know of any examples where a city councilman has gotten hold of your report and said, "Look, it shows that it doesn't make any difference. Let's cut the police force in half." And any examples of fights between the police force and people trying to cut their budget, interpreting your study as saying it doesn't make any difference?

MR. LEWIS:

There are a lot of them. I won't cite any particular one, but we had a mixed piece of fortune before we completed the report. As a matter of fact, we were still collecting data when a New York Times story was issued which said some things about what the study had found before we knew what they were. We had some feelings about it, but we didn't know what the data would show. That is a common problem in evaluation. But it was written quite well. It was David Burnham's article. He does things very well. He produced a lot of caveats, about use of unanalyzed raw data, preliminary police department impressions, etc., but he had a purpose in mind. He wrote his article at the time of the Mayoral elections in New York City. One of the things that all of the candidates were doing was out-promising each other in terms of the additional patrol officers they were going to add to the police force. He thought that that was, to put it in his terms, a crock. He'd heard about the study, so he went to Kansas City, talked to people in the police agency, decided that their views supported his views and wrote his piece, using some of the raw reported crime data which had not been analyzed at all. We were tracking reported crime data very carefully to assure the police that nothing really outrageous was happening, no one had stolen a beat or anything like that. That was the data he used.

That wasn't so bad. It disturbed us a bit, but I said it was mixed. One likes to be noticed as well. Anyway, what happened, which <u>was</u> really serious, was that the wire services and syndicated news services spread that story (because it was in the Sunday Times by David Burnham), all over the country. And they left out all the caveats. The worst example that I can remember of what happened was a compression into about three lines in Time Magazine, buried in the

middle of another article. I don't remember the exact words, but it was a little paragraph of two sentences or so which said in effect, "It is noted that a study done by the Police Foundation and the Kansas City Police Department showed that policing doesn't matter."

We thought that that was a very considerable compression of the information. Yes, George?

MR. KELLING:

Let me just add that there got to be an interesting conflict even earlier between the Kansas City Police Department and the evaluators. As far as the Kansas City Police Department was concerned, after eight months of the experiment, they were all done. They knew there were no differences. They didn't care to finish the experiment As far as they were concerned, the idea was, let's get on with it now.

Our stance at that point had to be, "Hey, hold it. There are other people we have to convince besides the Kansas City Police Department. For example, we have to convince an academic community that is going to review this thing. We need the year and we need all this data."

But as far as they were concerned, after eight months, let's call the New York Times and tell the world what we have found. In fact, they were heavily quoted in the article when it did come out; and they knew that it (routine patrol) didn't make any difference. They didn't care. We could do our fancy counting, but that wasn't terribly relevant for them, which developed an interesting conflict.

MR. LEWIS:

Knowledge in the viscera is always easier and quicker than knowledge in the head, I find.

The effect of the bowdlerization and the compression of things to a ridiculous point of distortion was that, though there are something over 20,000 copies of that summary report out, even people who have it can still misunderstand it because they remember those newsclips.

Anyway, city councils, the managers of many cities, have begun that way, as you suggest. Smart police officials have understood the report and use the report itself to say, "It says right here you can't use the report for this purpose. It doesn't address the manpower issue at all."

The careless acceptance of the notion that because some technique that is being used by service forces doesn't seem to make much difference--I am connecting that to whether you ought to cut its manpower or not--is very, very appealing; but it is naive. If you think how much undone order maintenance and crime control there is, it might suggest to you that if you knew something more effective to do, you'd want a lot more people. And that would be the issue. So that cutting the police is only one possible consequence to follow from such a study.

PARTICIPANT:

Bernard Greenberg, Stanford Research Institute. Now that Kansas City is on its second police chief since Clarence Kelley, what is the bottom line as a result of your experiment? What is left over? George said the police officers, patrol officers certainly knew what was happening. What is the residual in other words? Are they using, have they withdrawn patrol?

MR. LEWIS:

In the area in which they are testing new patrol strategies, they are reducing preventive patrol because they are using that time to do

the things that they are doing under the LEAA-funded program. But in other areas, as far as I know, I don't think any changes have occurred yet except that they don't emphasize rushing back out to the street the way they used to. They don't know yet what works better.

PARTICIPANT:

I am Jim McDavid from the Institute of Public Administration at Penn State. One of the points that Mr. Kelling mentioned is that people in the community weren't to be told that this was happening to them. Really two questions. Did at any point this knowledge get into the community, either from the New York Times story or from other sources, say, the police, telling people that they were responding to calls for service, that, yes, we are running this experiment and don't expect to see us on the streets. The second question is, what effects does this kind of withholding of information have on your ability to transfer the findings to other situations where obviously the community would have to know?

MR. LEWIS:

Let me take those in two pieces. The New York Times story came out after the experiment was over. We were still collecting the last survey data, but the experiment was over; so I don't know the answer to that. As far as we were concerned, it didn't matter by that time.

As George pointed out in the beginning, public announcements were made; but the area in which the experiment was to take place was not specified. The announcements were made in the terms that George described, well in advance of launching the experiment, to minimize people's attention to it. At the same time, they knew Clarence Kelley was trying something. The key to that is that Clarence Kelley had been there 12 years. He had the complete trust of his commission to which he reported and also of the citizens. If Clarence

said, "I am going to do something that is a good thing to do," everybody said, "Okay, that is probably a very good thing to do." That does raise questions about doing similar things in other places.

But let's be clear. It was never an idea that what we called reactive patrol would be a patrol strategy. There are people who note that in fire fighting we do pay for tactical readiness; and in between fires, the firemen paint toys. We do that. We are perfectly happy to do that. There are people who say, "I wonder if maybe the fire station sort of deployment for police, would make just as much sense." We can't prove that it doesn't, but that isn't what we had in mind; and nobody was ever going to recommend that another city adopt a reactive strategy. What we were going to say, "Look, if there is something you'd like to try that takes away part of the time available to your patrol force, take it out of routine preventive patrol and go ahead and try it. It won't do you any damage. If you can't think of anything else to do right now, think hard about all that you are doing--not just routine preventive patrol, but all of the things you are doing." It's safe to do that, and departments can do it. There was never going to be any pressure for someone to adopt reactive patrol as a strategy. Yes, George?

MR. KELLING:

In one business area, they did find out about the experiment through a patrol officer. Clarence Kelley went out and met with them along with the commander for that area and simply said, "Hey, this is important." The commander told them in very sincere terms what they were doing and why it was important, and Clarence told them. It turned out that, later on, that businessman's group wound up giving the major of that area an award for his innovativeness in policing. So it was a situation where Clarence Kelley had high credibility in the community, was well thought of in the community, had the support of his troops,

was able, when it did leak in the one area, to go out and explain it and explain it in ways which satisfied the business people.

PARTICIPANT:

How would that affect the interviews with the business people after the experiment--their perceptions of their own safety or victimizations or whatever kinds of measures were to be developed?

MR. KELLING:

Well, we didn't sort that particular beat out, but we simply found no differences in the three conditions essentially.

PARTICIPANT:

Would you call Chief Kelley an experimenting administrator, and has his success, or the success of this effort, permitted other people to become successful being experimenting administrators?

MR. LEWIS:

We certainly would call him a successful innovator. One of the problems that we had was that Chief Kelley was so committed to movement and change in his department, in policing generally, that he seized upon the notion of doing formal experimentation as an instrument for him to use. As a consequence, we were trying to do four at once. We learned you can't do that. You can do only one major experiment at a time.

As far as establishing or spreading the idea of inquiry and experimentation more broadly, I think it has helped. We are approached more and more--and I am sure LEAA is--by police administrators who haven't

done this before or who haven't done it in a very formal sense before, saying, "How can we get into this network? This is where leadership is."

This is a very fine thing. We are delighted. We hope it will grow. It does seem to be spreading, but we have got a very long way to go.

MR. GRANDY:

I think maybe we had better defer further questioning of this group. There is a lot of interest here and perhaps at our lunch break, those of you who have unanswered questions for these gentlemen can pursue them. We would like to thank you for a stimulating presentation.

Before we break for lunch, we need to say a few things about the organization of the working panels and the administrative arrangements. Mrs. Chelimsky will describe some of this.

MS. CHELIMSKY:

Well, while we are out at lunch, presumably all of this decor is going to be changed; we will come back and find rooms where there weren't rooms before, and other changes of this sort. We are going to break into the working panels directly after lunch.

As you know, Working Panel I, improving the interface between agency needs and evaluations, will be Chaired by Clifford Graves. Clifford Graves is now Assistant Chief Administrative Officer and Director of the Office of Management and Budget for San Diego County in California. Mr. Graves came to San Diego County after seven years in Washington, D. C., where he served, I think many of you know, two and a half years as Deputy Associate Director of the U.S. Office

of Management and Budget. Prior to that, he was Deputy Assistant Secretary for Community Planning and Management, Deputy General Manager of the Community Development Corporation, U.S. Department of Housing and Urban Development.

Panel II is going to be Chaired by Dr. Marcia Guttentag who is Visiting Professor of Social Ethics in the Psychology Department of Harvard. She was Director of the Harlem Research Center and Associate Professor of Psychology at the Graduate Center in the City University of New York. A social psychologist, she is co-editing a Handbook of Evaluative Research; and she serves as Research Consultant to the Office of Child Development. Panel II, of course, is the panel which looks at criteria of effectiveness.

Panel III, which looks at the utilization of evaluation, is Chaired by Blair Ewing who is Deputy Director of the National Institute of Law Enforcement and Criminal Justice. Before that he was part of the Planning and Management Section of the Law Enforcement Assistance Administration in the Department of Justice, and he was formerly a planner for HEW.

MR. GRANDY:

Let me say a word about room locations. Panels II and III will meet in this same area, where we are now, but the folding walls will divide this section off from the other two areas down here for those two panels. Panel I will meet in a large room in another part of the building. You will return to this same entrance of the building where members of our staff will be available to guide those of you through the building to the main briefing room where Panel I will meet. So those arrangements will be made; and when you come back, we will go directly into the working panel sessions.

LUNCHEON November 18, 1976

THE ROLE OF EVALUATION IN EXECUTIVE-BRANCH DECISION MAKING

JAMES M. H. GREGG, Assistant Administrator,
Office of Planning and Management,
Law Enforcement Assistance Administration

MR. GRANDY:

Ladies and gentlemen, we have a very noted luncheon speaker. It is our pleasure to welcome Jim Gregg, who is the Assistant Administrator at LEAA for Planning and Management. His topic today complements other parts of our program, particularly the speaker last night who spoke about the Congressional perspective. Jim today is going to talk about the Executive Branch perspective on the role of evaluation in decision-making.

He is a graduate of the Harvard Law School, has been a trial lawyer in Massachusetts and an attorney with the National Aeronautics and Space Administration. He has served with the Office of Management and Budget and has some experience there as a budget examiner and later as the Deputy Assistant Director for Program Coordination. He has had a diverse background in the public service. He was for a time the Assistant Director with the White House Special Action Office for Drug Abuse Prevention. For the last two years, he has been in his present position with LEAA.

We look forward to your remarks, Jim. It's a pleasure to welcome you.

MR. GREGG:

Thank you, Mr. Grandy. Ladies and gentlemen, colleagues and friends, let me begin my commending the Institute and MITRE for

convening this conference. As Government grows increasingly larger, more costly, but not notably more effective, a discussion of evaluation and its role in Government is very timely. When I was informed of the persons invited to this conference, I was impressed by the variety of backgrounds and perspectives that are represented here. It caused me to think again about this business of evaluation and made me wonder what criteria are usually used in inviting people to a conference on evaluation. Would it be people whose job description is evaluator? Would it be people in an office or division of evaluation? Would it be people who had recently been doing or managing evaluations? Perhaps all of these, but what exactly is evaluation? Why do we set it off with a job description? Why do we create an organizational entity to do it? Then my thoughts along these lines began to get a bit frivolous. Perhaps MITRE should hold a conference on the use of thinking by Federal agencies. Who would attend? People with a job description of "thinker"? People from the office of the division of thinking? Or perhaps who had recently been doing some thinking?

Idle thoughts, I guess, but it brought me back to the question. What is evaluation? and, Do we need it? Should we separate this function from the other thought and management processes of Government? I have a tentative answer. It's one which should please evaluators. The answer is, Yes and no. Yes, we need this field of specialization; but I believe that the great value that can be gained from evaluation will be lost if we compartmentalize the function. I want to suggest some ways of avoiding that danger.

As we all know, there is a great deal of motion and activity in the field of Governmental evaluation today. It reminds me of a sad but humorous story about a fighter pilot during the Second World War who never returned to his base in North Africa. Just before his

plane went down, a last radio message was received from him saying, "I am hopelessly lost, but I am making record time."

Those of us involved in Government and in evaluation of Government programs, I think, must empathize with that pilot who was moving in great haste, but not toward his objective. To the extent that we are not moving toward our objective, we need to pause and try to understand why we aren't. On other occasions, I have found it appropriate to be evangelical about evaluation, optimistic about its utility, enthusiastic about its potential, arguing that, in a period of increasing pressure on Government spending, an expanded role for evaluation is essential and that evaluation as a function should have greater identity and visibility. I do believe all of that.

However, with an audience of professionals, I must in all candor qualify this enthusiasm and say that while I foresee increased resources being devoted to evaluation, I also see serious limitations on the contributions that evaluation can make unless we begin to solve some of our broader problems of Governmental management and understand evaluation and its role in the context of the broader management problems and issues.

In the field of evaluation, there is much activity, motion and haste. We are making record time. However, until we begin to relate and integrate evaluation into the general thought and management processes of Government, and until we begin to greatly improve those processes, evaluation will realize little of its real potential to contribute to the attainment of our Governmental objectives.

In order to have further substantial contributions from evaluation, it is critical that we better understand the general Governmental management environment in which evaluators function and in which

evaluations contribute or fail to contribute to public objectives. To proceed without that understanding is almost certain to result in wastage of resources and frustrations on the part of all involved.

Therefore, let's consider first some of the unhappy features or characteristics of Government management as we find it today. Let's start with an examination of some of the most basic problems; then consider some of the common symptoms that stem from these problems.

It's my contention that the basic problem of Governmental management is quite simple in concept but extraordinarily difficult to solve, particularly in social program areas. The problem has three elements.

One, the failure of Government agencies and programs to set priorities.

Two, the failure to establish clear and reasonable objectives, and I emphasize the word reasonable.

Three, the failure to hold Governmental managers accountable for results.

Where these three problems exist, and they exist in many social programs, it usually reflects a failure of leadership on the part of both the Congress and the Executive Branch of the Government. Time doesn't permit me to fully develop for you the exact specifications of each of these failures. You are perfectly familiar with them anyway. However, I will mention some of the common symptoms or problems that do flow from the basic failures.

You will find many legislative enactments that call upon Federal agencies to accomplish miracles, often with totally inadequate resources provided, which, I suppose, is part of the miracle to be performed. You will find legislation in Federal programs with long lists of goals and objectives, often very general in nature and far too numerous to be subject to any reasonable degree of top management supervision or control. And not infrequently, the vague objectives mandated are contradictory or conflicting. You will find greater interest in Federal agencies in obtaining and spending larger budgets than in getting results. You will find confusion among agency staff as to what is truly imparted, both in programmatic terms and in terms of their own accountability. This often translates into poor morale and low productivity of staff.

While you may find some interest in efficiency, you often find little interest in effectiveness. And you will find little interest in systematic management processes.

I could go on, but let me stop here because I believe in attempting to deal with the symptoms of poor management processes, we are learning a few lessons that can lead to some solutions to the basic problem. Of course, the ideal solution is to have a Congress return next year that only legislates after careful consideration of the issues, which mandates a limited number of high priorities with clear and reasonable objectives included in its legislative enactments, and have an Executive Branch with managers whose principal accountability is for reaching the objectives that have been defined. I don't expect that to happen.

There will continue to be too little deliberation, goals will be cosmic, objectives will be multiple and perhaps conflicting, there will be too much money appropriated and intense pressure to spend it;

and finally, when the absurdity of it all is evident, there will be a demand for evaluation and great finger-pointing when evaluation reveals the inevitable failures.

Can we find a better way to do the public's business that is more in the public's interest? I believe so. It will require a joint effort between the Congress and the Executive Branch. It will take a long time because it will require new ways of dealing with problems and Governmental responses to them. It will take the talent and heavy involvement of people with your skills to make it work.

To make it happen, it seems to me, three things must occur. First, strategic planning for results must become the rule rather than the exception in Federal domestic programs. The whole process of strategic planning and programming must be undertaken routinely by domestic agencies; and the most critical component of this process is a tough, realistic approach to priorities. As long as the Congress and the Executive Branch pretend that the Federal Government can solve all problems, we will solve few if any problems.

I would like to add here that evaluators and the results of evaluations must be involved in the priority-setting process. This involvement could help us avoid undertaking the impossible which we frequently have done in the past.

Secondly, new program development must be a more deliberative and systematic process than in the past. If this means we move more slowly to address our social problems, so be it. It is results that we want, not just motion and record times. Too often in the past, we have attempted to develop new social programs through funding so-called demonstrations. Too frequently, these have not been carefully phased developments of program concepts into new projects or programs

having specific design and performance specifications. Too often such demonstrations have represented only rough attempts to generate new ideas and new programs by providing funds for a great variety of projects meeting highly general criteria. Programs developed in this fashion are difficult to evaluate. Often evaluation is not even considered until program implementation is well underway or completed.

Project and program development must be done more slowly and with greater care. There must be careful design of projects or programs at the beginning, with performance measures specified. There must be limited testing of the project and program concept. Some redesign should be anticipated after the testing phase. Only then should there be broad demonstration with predesigned evaluation as part of the demonstration.

All this does take more time, but when we proceed in this fashion we know what we are doing. We know what performance can be expected from a project or program and at what cost. Even when we fail, we are more likely to understand why.

The third and final requirement to make it happen is to establish accountability of Government managers for results. Need I tell you that it often seems that there is accountability for everything but results. There is accountability for fidelity to a policy line even when the policy is vague or ill defined. There is accountability for good public and Congressional relations. There is accountability for spending one's money promptly. There is accountability for assuring compliance with a thousand and one Federal laws and regulations and so on.

But to make Government work, we must establish accountability of Government managers for program performance and program results. We

cannot do that until we have limited and realistic priorities, clear and specific objectives and resources reasonably commensurate with the objectives to be achieved.

So we come full circle back to priorities and objectives, and the question remains, "Can we move toward a more deliberative and perhaps more experimental approach to Governmental programming with accountability for results?" I believe the answer lies with us and what we can contribute to such a movement. Also I believe there are economic realities and strong political currents that will carry us in that direction. We have frustration of citizens over the high costs and poor results of Government. We have greater political sensitivity to the need for reform of Government management. We have fiscal pressures that remind us that we cannot afford vast waste in Government. And not least important, we have increasingly the techniques for programming in a more rational and even experimental mode. Increasingly, professional program development and evaluative skills and techniques will be brought to bear on Governmental programming. You and I are on the frontier of this development. We are still somewhat pioneers in the application of these techniques to social programs and problems. Twenty years from now, I am sure we will look back and marvel at how primitive our techniques were for good program development and program evaluation.

But our mission today is not to speculate on how we may appear to more sophisticated generations, but rather to get the change we need underway; and by our activities in this and the next decade, accelerate our progress toward more effective delivery of Governmental services.

I appreciate the honor of having been invited today and wish you all well in your deliberations. Thank you very much.

MR. GRANDY:

Do you have any questions you might like to ask?

PARTICIPANT:

I am Seymour Brandwein, Department of Labor. I recently met with a very insightful British analyst who was wondering about the spread of the ideology that Government is ineffective here. He compared the American and British systems and came up with his observation that the British system of starting programs small with a carefully developed design assured that the programs remained small and that the design remained inflexible. He praised the American will to do and to learn while doing.

Without saying either is wholly right, isn't there a danger of killing the animal while trying to cure some of its diseases?

MR. GREGG:

I think there is that danger. I think you have to leave options open. I know in our own agency, we have been discussing program development rather intensely just recently and how it should be done, and there is a strong feeling, and I agree with it, that you should not put all of your chips on this approach. Sometimes, in fact, you don't really have the political option to do that even if you would like to. It's probably not wise anyway.

I suppose in response to the British gentleman's observation, I am not sure that we can afford to continue the "learn by doing" approach to Governmental programming. Perhaps it's a luxury that we have been able to afford in the past, and I am not convinced that the benefits of that approach have been great. Even if they were, I am not sure that we can afford that kind of approach in the future.

PARTICIPANT:

I'd like to add first to your last observation. Isn't it possible that what the British gentleman was referring to is a different phenomenon altogether? For example, speaking for myself, the only institutions that I know about that work well, that appear to be smooth externally, are monolithic. I think the British commentator knew the weaknesses of his own system and extolled the virtues of ours, just as we do, based on our own deep knowledge of our own problems and our ignorance of other people's.

The question I was going to ask is how would you propose to cure one of the maladies that you referred to? It's very difficult to mobilize political and other energies toward the redistribution of resources--everybody pays attention to that--without overpromising. It seems that vagueness and overpromising are a part of the process of getting legislation passed. Is there a way around that?

MR. GREGG:

I don't think there is an easy way around it. That is why I indicated in my talk that I think we are in for a very long haul in getting this kind of change. I think it is an attitudinal change. There are all kinds of political pressures involved in accounting for the reason we do things the way we currently do them. I think both within the Executive Branch and Congress, gradually we are going to have to realize we cannot do business in the way we have been doing it. There are techniques available to us for doing it in a more rational and economical way and probably getting greater results. I hope we can find a way through which the particular political needs of Congressmen can be met in the process. I must confess it's a very challenging problem, and I think it's going to take a long time.

MR. GRANDY:

Thank you again very much for coming.

BANQUET ADDRESS November 18, 1976

EVALUATION FINDINGS: THE CASE FOR MARKETING

CHARLES R. WORK, President,
The Distirct of Columbia Bar

MR. MASON:

While you are finishing up your dessert and coffee, I have the privilege of introducing tonight's speaker. I am Bill Mason[39] from The MITRE Corporation, and I am honored to introduce Mr. Charles Work who is going to speak to us for a few minutes tonight. Most of you know that Mr. Work was a Deputy Administrator at LEAA before he went into private law practice, and he was with the U. S. Attorney's Office in the District of Columbia before that.

He is now President of the D. C. Bar Association. We have asked Chuck to tell some of his experiences while he was the Deputy Administrator at LEAA where he was involved in starting a number of evaluative efforts and has had an interesting set of experiences while trying to incorporate evaluative techniques in the mainstream of the decision-making process at LEAA.

MR. WORK:

I don't know how to respond to that introduction. When they just read my resume, I say, well, that is the story that my mother likes to hear and my father doesn't believe. He is delighted that I am in the private sector. He thinks I am earning a living for the first time.

---

[39] William F. Mason, Technical Director, The MITRE Corporation, METREK Division.

I am pleased and honored to appear in front of this group tonight. I am always somewhat intimidated, however, when I appear in front of a group of experts, and I know from my first-hand experience that there are a great many "experts" in the evaluation field. Talking to experts always reminds me of my very first appearance in a courtroom in a major metropolitan city--it was Washington. In that environment, the young prosecutor is expected to walk into the courtroom his second day in the office (he gets one day to get acclimated) and he is handed a file folder and is expected to try the case. Well, in my day, the manila folder was a major management revolution, so the file folder was a piece of 8½ x 11 paper that was folded over as I just folded my speech. If you were lucky, the system worked well enough so that the "folder" reached you in the courtroom in time for trial. If you were luckier still, the witnesses were also there. My second case, my first day, was rather unusual. I was prepared for a petty larceny-shoplifting or a drug case. I was not prepared for a Murphy game (or con game) case. You see, the police hardly ever catch the perpetrators of a Murphy game.

This Murphy game involved two con artists, and usually, working this game, they pick on a young tourist. It works something like this. The young tourist standing out there on the street obviously is looking for trouble.

One of them approaches the young tourist and says, "Look, are you looking for a good time tonight?" And they have a conversation about it. The second con artist comes up and all three of them have a conversation. They say, "Well, let's go out and get in trouble tonight." The first con artist says to the second con artist, "Look, if we get in trouble we have to be able to rely on each other, trust each other."

These two con artists, of course, are pretending like they don't know each other. The first con artist says to the second con artist, "How do I know I can trust you?" The second con artist says, "Look, I am going to give you my wallet, and you and our young friend tourist here--you walk around the block. If you come back, I know that I can trust you." Of course, then the first con artist hands the second con artist the wallet, and the tourist and he walk around the block and come back. It's the first con artist's wallet. Then it's the tourist's turn. Of course, the tourist stands there and the two con artists walk around the block with his wallet and never come back.

As you might surmise from those complicated facts, being nervous and upset and merely anticipating a shoplifting case, I was not ready for this case. I couldn't even figure out what witness to put on. The judge understood my puzzlement and leaned down over the bench and said so everyone could hear, "It's all right, Mr. Work. You may not know what is going on here, but all the rest of us do."

That's the way I feel talking to evaluators.

I am not an expert in evaluation. I am not methodologically sophisticated. However, you might say that I am an expert in receiving evaluations. I have received all kinds of evaluations. I have received quick and dirty evaluations. I have received slow and clean evaluations. I have received one-page evaluations. I have received thousand-page evaluations. I have held my breath while an evaluation was going on of a project that I really believe in and had worked hard on. I have despaired about evaluations that glossed over problems that I knew existed. I have

helped to plan evaluations after the program was over, while the program was running, while the program was being developed, and so on. I have even run a project at LEAA which attempted to find 1001 promising projects; we called it Project Scheherezade. I have to tell you a little bit about that.

I was often asked while I was at LEAA, "Tell me about one program at LEAA that has been a success." If the person were really well informed, he knew that we had funded 80,000 projects and so he would make it an even tougher question and say, "Give me one program out of 80,000 that has been a success." So I told the staff that I would like to see one-page summaries for a thousand successful LEAA projects. Someone picked up on the 1,000, made it 1,001 and called it Project Scheherezade. There is such a document. We didn't find a thousand, but we found 600. I was secretly pleased--600 out of 80,000 for a Government program--didn't seem to me to be so bad.

The fact that this symposium is being held is, in my mind, a recognition of the growing importance of the field of evaluation, if that is what you want to call it now. I believe that it is safe to predict, and I understand that you have heard from a number of speakers, that this field will become more important during the next few years. In my mind, it's not that you all haven't done this kind of work before. I frankly don't know the difference between research and evaluation or evaluation and program analysis or whatever you wanted to call it a few years ago. But I would suggest that merely by calling this process--whatever it is--evaluation, we are indicating that we are trying to make this kind of work more relevant to the policy and decision-making process. And in my mind, there is clearly a need for this kind of relevance.

It is clearly in demand. In my view, the public is demanding it because of a seeming decrease in resources, a growing lack of patience with bureaucracy and no results. And so are, believe it or not, some state and local Governments; and so is Congress in its own simplistic kind of way. Of course, saying that it is important and that it is growing in its importance doesn't make it easy to do.

The interviews that were conducted in advance of this symposium were sent to me, and I was pleased to be able to read them. But only a cursory review of those will reveal that this is an exceptionally complex and difficult area. A conference of those persons involved in the evaluation process in different Federal agencies is a tremendously ambitious idea. It's not, I'm sure, lost on you that evaluation in a place like the Patent Office is different from an evaluation in the Census Bureau, ~~is different from an evaluation in the Census Bureau,~~ is different from an evaluation in LEAA; and so trying to find a thread, trying to find an abstract level to which we can all respond is exceptionally difficult, even given the obvious fact that there is an increasing demand for evaluation and evaluation results.

But there <u>are</u> some common threads.

We all seem to have a methodology problem. There is an exceptionally large gap between those of you who are methodologically sophisticated and those of us who know nothing about methodology.

We all complain about data. There is a lack of it in general, and most of it seems to be of dubious reliability.

We cannot decide what the role of the evaluator ought to be--whether he is the independent, reliable critic or the person helping the manager "fine-tune" the program.

Finally, we all seem to be haunted by the question of whether, even though evaluation has become more visible, even though there are Assistant Secretaries with evaluation in their title, whether or not evaluation really does make a difference.

I'd like to propose that we look at these problems and a few others somewhat differently. I'd like to propose that we look at them backwards--review in our own minds a finding that we are familiar with from the evaluations that we know and ask ourselves what happened to them and why. Ask ourselves what were the results that came about as a result of those findings, what difference did they make, what kind of changes did they make and for what reason.

I have chosen to call this process examining evaluation from the marketing dimension. It is my premise that evaluations should be useful and that they should make a difference. It is my thesis that if this marketing dimension is considered, that there is a much greater chance that the evaluations that we do will make a difference and will have an effect. It is also my view that this marketing dimension has to be considered at the beginning of the evaluation, not at the end of the evaluation; and if this marketing dimension is examined at the beginning of the evaluation, that it may make a difference in the program itself. It <u>will</u> make a difference in what the program development actually entails. The question is, how do you sell or market the results of the evaluation. What

will make the results relevant and useful? It is not, in my mind, just salesmanship; although salesmanship of these findings is a dimension of it. The marketing dimension in my mind provides an extremely important critical eye. It will result in a higher quality product.

This approach also helps us focus on some of what I consider to be certain important generic problems in evaluation. And, as I look at the evaluation process, some of these thoughts that I am about to advance are confirmed.

Clearly, one of the most important problems and something that has to be considered in developing this marketing dimension is the problem of the communication gap between the evaluator himself, the evaluation team, and the decisionmaker or policymaker. It is really an exceptionally difficult problem in any field. But it's particularly difficult in the criminal justice field because the criminal justice practitioner and the evaluator are so far apart in their experience and their outlook. I often think how much easier it would be to be in the field of medicine or health or even in the field of education where there is some similarity in training and some other common ground. What does a criminal justice researcher have to say to a police chief who did not graduate from high school? The only practitioners in the criminal justice field that have had any degree of higher education are the lawyers; and, of course, the lawyers are so singularly insular and isolated in their outlook that they cannot understand or comprehend anything that isn't written down in a legal case book. Only in the law would the so-called Brandeis brief be a revolution.

Another way of stating the same problem that may be more relevant to those of you who are working within an agency and are working with persons who are developing programs within an agency, is as follows: A policy-maker may listen to the findings and then decide that the question he wanted to ask was entirely different from the one that you have been working on for 10 or 12 months. I had that very thing happen to me when I was the Chief Prosecutor for Local Crime in the District of Columbia. I had asked our researchers (and I was very proud of the fact that we had a research team--this was back in the late '60's) how many cases were being dismissed because of lack of witness cooperation. I knew it was a problem. They literally worked thousands of hours, and it was months later that our researchers were able to return to me full of pride joyfully saying, "Mr. Work, 38 percent of all of your cases are dismissed because of lack of witness cooperation." I looked at that figure. I was ecstatic initially. I had a figure. Then I said to myself, I knew it was a problem all along. What good is it to me to know that it is 38 percent of my problem?

At any rate, after I figured out what was wrong with merely knowing the figure, we embarked on yet another evaluation because we had to know why the 38 percent of those cases were dismissed for lack of witness cooperation. The result of my asking the question of why 38 percent were dismissed is a book; it is entitled Witness Cooperation by Frank J. Cannavale, Jr. and William D. Falcon. We ended up doing a household survey of uncooperative witnesses, and I believe it is an important piece of work. I now cannot fail in any evaluation context to say to myself, well, when I get what I am asking for, what will I do with it? I learned that the hard way. That, of course, did not happen to me just with the witness

cooperation question, but it happened to me with all those percentages that were returned in first or initial kinds of evaluation reports. The figure itself, the first question itself, just may not be very helpful and may not be very useful.

Another generic problem highlighted by considering the marketing dimension is the problem of timeliness. The policy-maker and decision-maker will always tell the evaluator that he has to have that evaluation done tomorrow or at the very least, for the next budget cycle (which is almost tomorrow), because the decision-makers and policy-makers are always in budget cycles.

But I found in my experience that time and time again, the evaluators fail to make the deadlines for those budget cycles. And there nevertheless seems to be another budget cycle that the evaluation will fit into eventually. My reaction to the problem of budget cycles and evaluation is that evaluators ought to take their time and forget about the budget cycle because there will always be another one that they can fit that evaluation into.

Well, of course, there are other problems that can be characterized as marketing problems, but I'd like to turn to just a few ideas I have about overcoming them.

It is certainly a truism that many evaluations fail because of definition problems. I think that the person who begins an evaluation of a program that is ongoing, rather than being involved in planning the evaluation when the program itself is planned, is under a very severe handicap. It's incumbent upon him to define what is being developed, what is being evaluated, what ought to be evaluated, who his audience is, what they want to know. It

seems to me that the policy-maker cannot be counted on to ask those questions himself, especially if he is already into the program. If those definitions cannot be agreed upon or if the questions cannot be answered, if there isn't a methodology available or the data available to answer those questions, then I think it's incumbent upon you to say initially, "Look, Mr. Policy-maker, Mr. Decision-Maker, we cannot answer that question."

If you are in on the program development stage, you are going to have a much greater chance to change that and to identify those questions earlier. It is my experience that you can get the program objectives, the program definition changed when you can show that you really can't produce a useful evaluation any other way. In short, there is no excuse for today's policy-maker in the Federal Government not involving the evaluators in the program development stage.

Of course, coming from the policy-making and decision-making side, I feel strongly that evaluations in their ongoing stages ought to adopt a no-surprises outlook. I realize that that might be disputed in this particular audience, but I think it's extraordinarily important to keep the program manager briefed as you go along. I think the confidence of the program manager is an important thing for the evaluators to have. But more important than that, I think the program manager isn't doing his job unless he is asking the evaluators what they are finding as the program goes along. I certainly wouldn't be running a program in the Federal Government today that was being evaluated without finding out what the evaluators were finding out. Even if the evaluators were worried about their credibility and independence, I'd still be grilling them and trying to find out what they were learning.

Next, as Dick Linster of LEAA has said, I think that consideration ought to be given to using a similar methodology to evaluate different programs within the same office whenever possible. I think that the idea of putting three or four evaluations together is very exciting. I think that the idea of acquainting program managers with certain methodologies and getting them to understand some of these methodologies would also bear fruit. It will in the long run pay, not only programmatic dividends, but pay dividends in terms of the relationships between the evaluators, the decision-makers and the policy-makers. I think that we ought to be striving for some symmetry. I think it would help this marketing dimension substantially.

Cost/benefit analysis is a methodology that might well be more widely applied in order to achieve some symmetry. It has one important advantage, and that is that people like me understand it. If you can put a dollar sign on something, I can understand what you are saying. It is something I can follow. Even though it may not be relevant to most of the things you are doing, there may be some dimension of your evaluations in which cost/benefit analysis would be helpful. I think that it ought to be applied, even if it is just to that relatively narrow segment of what you are doing. My views are colored by a particularly successful cost/benefit analysis of the LEAA Comprehensive Data System Program done by the Institute for Law and Social Research in Washington, D. C. There were a number of assumptions in the program that econometricians were able to destroy. One was the notion that the program was going to get cheaper. It turned out that it was going to get much, much more expensive. LEAA just didn't understand that as they were developing the program.

One idea that I have seen used with great success in evaluations is what some people call "user groups". I sort of stumbled on it, but I think it could be used more successfully than it is presently being used by evaluators, and I don't think it matters whether you are in criminal justice, health or whatever. The user group idea involves calling together the people from all around the country that would produce or use these research findings and get them to talk about them. Bringing them together seems to relieve some of the anxieties. They say to themselves, "It's not just me that has these problems. This other place, this other jurisdiction also has the same difficulties." Bringing them together and hearing them say, if you will, that I've got this problem and I've got that problem seems to me to have an energizing effect. It seems to break down some of the resistance to looking at something objectively, admitting something isn't going quite as well as it ought to go. They don't feel so alone. They don't feel that it's them against the evaluators, and I think that is an important marketing insight.

Finally, I wish to make a proposal following on a number of points that I made here tonight and really based in part on the success of this conference and the problems that it has revealed. It would seem to me that even though there are many disparate problems, even though different kinds of agencies have different kinds of evaluations, that our common objectives would be served by bringing together at the Assistant Secretary level a group that would try to foster and develop and compare information on evaluations that are ongoing throughout the Federal Government. It would seem to me that such a structure would have a working group level, and the working group level would meet more often and would be responsible for the development of the program, of the agendas, and really trying to look systematically at the exchange of this kind of

information. I think that this exchange of information would be very helpful. It would help to market evaluation and it would carry forward the spirit of this particular meeting. I think that the notion that we have just begun, that this field is just beginning to surface is correct. And I think that in the process of surfacing it, we can speed it along, we can develop it if we will adopt some of these ideas that will enhance the interchange of various thoughts about this difficult and complex subject matter.

I have enjoyed being with you tonight. I want to thank you very much and wish you the best of luck in your endeavors.

I. WORKING PANEL III: IMPROVING THE UTILIZATION OF EVALUATION FINDINGS

CHAIRMAN: BLAIR G. EWING, Acting Deputy Director, National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration

MS. CHELIMSKY:

Can we convene now for the final reports of the panels? We have had, as you know, three groups meditating and reflecting on various aspects of our evaluation problems: Working Panels I, II and III. I had naturally thought we would start with Panel I and go through to Panel III, but given the kinds of issues which were, it seems, actually examined by the panels, it now seems more logical to reverse the order and start with Panel III. In this way, we can examine what the various panels have had to say, first, about users and conditions for use, second, about evaluation criteria and their substance, and finally, about evaluator/agency working relationships. Do you want to start then, Blair?

MR. EWING:

We are the panel on improving the utilization of evaluation findings. In my introduction to panel discussions yesterday afternoon, I said that it seemed to me that there were multiple uses of evaluation findings, ranging from program development to resource allocation, to killing programs, to covering various parts of administrators' anatomies, to planning, to the development of further research and further evaluation, to budget justification, et cetera. There were also a very large number of audiences within agencies for evaluation findings and these audiences could range from the program managers themselves (whose programs are being evaluated) to the planners, researchers and evaluators in the agency, to the top

management of that agency. I also talked briefly about some of the conditions that I saw as being essential for the use of evaluation results by agencies and then our panel began an enormously lively and spirited discussion which reflected quite a sum of experience and many discordant viewpoints.

We did not have time to address adequately all the topics on our agenda, perhaps because of some lack of consensus among us (although there was, in fact, some agreement), or perhaps because we were so preoccupied with those we did discuss in depth. We focused on three major aspects of evaluation use and usability:

(1) The user or the audience for evaluation findings;

(2) The kinds of information needed by that user, that audience; and

(3) The conditions which stimulate or impede the use of evaluation findings by agencies.

First, Evaluation Users. We began by examining the question of who uses evaluation findings, and decided that although there are many potential users within a given agency, the primary audience would depend on who needed the evaluation, and on the evaluation's character and scope. Given that users are pluralistic (decision-makers sit at different levels) and that there are many possible conflicts among the information needs of different users, the panel agreed generally that evaluation must at least begin by addressing the needs of the person who asked for the evaluation. The character and scope of the evaluation are also important in determining the audience for the findings, in the sense that an evaluation of a small program's efficiency might have as its major users the program's manager and the agency budget officer; whereas the users of a comprehensive, large-scale evaluation of the effectiveness of an important agency program would be the agency's policy-makers, and then--in widening circles--the research community, OMB, GAO, the Congress, the press, the public.

It was pointed out also, and largely agreed, that the audience for evaluation findings which concerned programs in the field, could not be limited to people at the Federal level since those findings needed to be implemented by state and local government people and by the institutional practitioners (e.g., teachers, policemen, nurses, etc.) whose work had been evaluated and whose efforts and good will would be needed to improve the program. Panel members felt that the Federal role in this area was to build knowledge and that efforts are presently lacking to improve the local ability to rank priorities or compare rationally among local programs as to effectiveness and cost. There was some consensus that--in the words of one participant--"When the Federal Government sponsors an evaluation, that evaluation gets designed on the basis of assumptions made by the Federal agency about what is of interest to locals. There is little or no participation by locals in the evaluation design. When the results come in, the Federal agency itself has difficulty in understanding what they may mean (either to the Federal Government or to state and local governments) and it has no strategy for communicating what they might mean to the local practitioners who are intimately concerned." Finally, the point was made that, where federal initiatives at the local level are concerned, there does not seem to be much point in doing evaluations of "demonstration" programs unless there is some commitment on the part of local people to institutionalize. In effect, if locals don't intend to continue a project, their need for evaluation findings would appear to be somewhat diminished. As one panel member put it, "The Federal Government has little leverage to ensure improvement at the local level, no matter how good the evaluation."

Possible conflicts among the needs of evaluation users was discussed at length. We recapitulated some of the Agency Perspectives Panel discussion by examining the public interest versus the agency

interest, Congressional and OMB oversight needs versus agency needs; we contrasted the Federal policy-maker with the local practitioner or implementer, and the Executive Branch generally, with the Legislative. Professor Martinson told us that "the fundamental function of evaluation, like other forms of social science, is to enlighten the public as to whether or not the agencies to which the public pays taxes is using that money properly." He felt that if that interfered with what he called "purely symbolic activity snugly ensconced in an agency," well, then so much the better. Most of the rest of the panel, however, felt that our panel was dealing with agency use of evaluation findings and that the users we should consider, therefore, had to be primarily the agency managers who had asked for the evaluation and/or needed the information it could furnish. One of our panel members (who represented a Federal agency) made the point that Executive Branch policy-makers cannot change important agency policy without Congressional assent; yet often, an effort to change agency policy because of feasibility, or cost/effectiveness considerations, runs up against Congressional attention to special, powerful constituent groups. Therefore, it is wise as well, to build in, early on, both Congressional knowledge and use of agency evaluation.

Second: User Information Needs. It seems a natural assumption that Federal agencies would be more likely to use evaluation findings which produce information needed by agency managers. From there, it seems only a small step to ask the decision-maker who called for the evaluation what he expects from it, what it is he needs to know. Our panel members did indeed agree that the question of whether or not evaluation is used by agencies does depend in large measure on whether the right questions have been asked. The problem is that it is often very difficult to find out what these "right questions" are, especially in evaluations of complex programs.

To begin with, all questions are not answerable, so the first problem is to find the three (or so) questions which can feasibly be addressed by the evaluation and which are important to the decision-maker. But, as one panel member pointed out, many decision-makers do not themselves always know the "right" questions to ask, and here the panel felt it might be a useful learning experience for policy people to be involved in evaluation planning. "The real issue" said an evaluator member of our panel "is training managerial people to understand the limits of evaluation," how it can be used, what can be asked of it.

Here we had a split in our panel. Some people felt that the way to find out the right questions was through direct interaction between evaluators and agency managers, that the latter don't need to understand the limits of evaluation. They pointed out that perhaps decision-makers do not need to ask questions better because there is too much lack of consensus in social program areas. "What decision-makers are really interested in," said one panel member (a decision-maker himself), "is in keeping the system operating and stable, in not letting the temperature go too high or too low. He doesn't want to transgress boundaries, he wants to know when it's too hot and when it's too cold, and whether the thermostat moves quicker in a heating or a cooling system. He wants to know whether there is money waste, and he wants to know whether there is any visible achievement, or any visible failure to achieve. Those are the 'right questions' for him."

Other panel members pointed out that the "right questions" depended upon the type of evaluation envisaged, that many agencies use evaluation almost exclusively as a management tool and that questions of program achievement and effectiveness could rarely be

addressed by such evaluations. So that, to promote use and avoid disappointment, it becomes very important that decision-makers understand what questions can be asked of a particular evaluation and how this information obtained can then be used. Some members suggested that participation in evaluation planning might be a useful exercise for allowing agency managers to familiarize themselves with the possibilities and limitations of various evaluation strategies.

My own feeling is that an important problem in establishing what questions to ask is that it is very rare (at least in my experience) for managers to call for evaluation in order to improve planning and decision-making. The questions they ask, and what they want to know, is a function of why they asked for the evaluation in the first place. Usually they ask for evaluation:

- when they are stuck with a program they mistrust and want to cover themselves;
- when the program is in an enemy's province (evaluation is here used as an assassination instrument);
- when they don't understand a program and want enlightenment; and finally,
- when Congress says they <u>have</u> to evaluate.

This may well be because evaluators have not communicated well enough with managers or because the other uses of evaluation have not yet trickled up. These ideas, then, do support the need for more understanding of evaluation among decision-makers, or at least some liaison mechanism, some bridge between evaluators and agency decision-makers.

<u>Third: The Conditions Which Make for Use</u>. I began my exhortation to the panel by listing five conditions for use, with which a good many members of my panel and members of the audience disagreed, but that didn't shake me any. I still believe these conditions are essential conditions.

I think in order for somebody to use evaluation results, particularly a manager or decision-maker--and this is, I think, all the more true the higher you go in the management hierarchy--that the information to be presented from an evaluation has to be reliable. It also has to be brief. It has to be timely--that is, the information has to be presented at a time when the manager can use it for a decision. It has to be comprensible, no jargon and careful writing. That was, I gather, Chuck Work's prime point last night. It has to be at least to some degree conclusive on some of the questions, if not all of the questions, raised in the first place.

The issue of what kind of structure or organization best promotes the use of evaluation findings gave rise to a great deal of fairly acerbic discussion. One panel member wanted us to stipulate that, for evaluation findings to reach policy-makers, there needs to be a centralized evaluation office in the agency, dedicated exclusively to evaluation (i.e., without responsibility for programs) and possessing close and constant access to decision-makers. This was opposed on several grounds:

- that agencies differ in terms of where the power is and where the needs are;
- that people low down in the bureaucratic hierarchy need (and should get) evaluation help, too; and finally,
- that such an organization would ensure only that the basic purpose of evaluation (i.e., public enlightenment) would be foiled because evaluation offices in agencies distort evaluation to suit the purposes of the agencies and the capabilities of the evaluation offices.

It was pointed out also in our panel that no evaluations are ever really conclusive, and that reducing jargon doesn't ensure the

comprehensibility of evaluation findings if decision-makers do not understand evaluation and have not been successfully reached by the evaluators. (Again we came back to the need for a bridge, a mediator between the evaluation and its agency user.) It was also stated that relevance (i.e., again, the "right" questions) and timeliness were more important then conclusiveness. Said one panel member, "Usability is not synonymous with rigor. Some poor evaluations have been used very constructively." Further, there is even some conflict between rigor and use, at least in some cases, because the more an evaluation resembles an experimental design, of course, the more reliable the results will become, but the less likely the evaluation is to be brief and timely and comprehensible.

Our panel did reach some conclusions, and let me state those as I understand them. They weren't shared by everybody, but I think they represent some conclusions by at least a majority of those who stuck with us.

I think those conclusions were that in order for evaluation to be used, the very first criterion is that, before you ever start on an experiment or a program or whatever you want to call it, you have got to find yourself a user, somebody who wants some information. If you don't do that, then you won't ever find anybody who is really going to use it in the end. That seems sort of like a simple proposition, but I think it's one that fairly frequently gets overlooked. As eagerly as users may be sought, they are not often found. Joe Wholey of the Urban Institute said that he has spent a number of years searching Federal agencies for people who considered themselves decision-makers or users and had rarely found any. Since he has done a lot of work for our agency, I assume that reflects on us as well as others.

But beyond not being able to find a user and beyond the necessity of finding one, comes the question of what it is that might be done to improve utilization once you have found a user (if you can find one). That question is a very complex one for which we didn't really find any prescriptive answer which would suit every case, but which involved, among other things, evaluators recognizing that it's essential that they should not merely sit by passively, but seek out opportunities to talk with people in policymaking positions to insist on a role for evaluation, at least insofar as they really believe that that is a proper kind of activity; in short, they should be aggressive about selling their wares. Now, that doesn't mean being aggressive about selling their wares when there is no real need for evaluation, but it does mean that in some respects, evaluators have to understand that in order for anybody to want to have evaluators around and to do evaluations, they have to be useful evaluators which means they have to produce things that people want. Particularly, they have to respond in many cases to short-term questions.

What you have to do, we concluded, is to have a kind of mix; and you have to be able to sell a kind of mix in your agency--a mix of short-term analyses and longer-term inquiries and some assessment and some disciplined judgment and also some evaluations and perhaps some research. That kind of mix is not very satisfactory from the point of view of people who are researchers by training and by inclination. But it is probably an essential kind of activity if the evaluation function is to survive at all as an evaluation function.

We did not, I think, reach a great many other conclusions in particular on which everybody agreed, but we did, I think, conclude that it is essential that there be much greater clarity about what it is that is promised in advance by evaluators about evaluation.

There have to be bargains made and negotiations undertaken at many tables. The more complex the Federal program, the more tables to which people must go. Which is to say, if there is a program that involves state and local governments as well as the Federal Government in direct program activity, then there have to be bargains struck all the way through about what the evaluation will do, whom it will serve, what questions it will answer, what it will produce and what kinds of results are expected at what cost. Those kinds of bargains then have to be also taken to another table, which in the case of Federal agencies includes OMB, and also the Congress. There are many bargains to be struck about evaluation. The clearer those can be in advance, the better off the evaluator is likely to be because then he or she will know what it is that it is necessary to produce in the end, and the more probable it will be that evaluations can be relevant, timely, understood and used. Thank you.

REPORTS OF THE THREE WORKING PANEL CHAIRPERSONS (CONTINUED)

II. WORKING PANEL II: IMPROVING THE DEFINITION OF EVALUATION CRITERIA

CHAIRWOMAN: MARCIA GUTTENTAG, Director,
Social Development Project
Harvard Graduate School
of Education

MS. CHELIMSKY:

Let's hold any questions and go right on since there is not much time left. Marcia, would you like to tell us what Panel II found?

MS. GUTTENTAG:

Our task was to discuss improving the definition of evaluation criteria. As someone left our meeting at the very end, he said to me, "Is it really possible to disagree with everything that everyone else has said here?"

With that as a caveat, whatever I present is necessarily distorted and shaped to make it sound as though there is a little consensus around what was said. I have six points to make which summarize our discussion, then a couple of conclusions, and a pessimistic and an optimistic note on which to end.

Point Number 1. We began by discussing effectiveness and efficiency criteria and decided early on that these were only two among many and that it was important not to use them as the sole criteria, this for several reasons.

First, that there are an enormous number of different means and different ways of operationalizing each of these constructs

and that, because of this, any single choice in operationalization is bound to exclude others.

Second, that they often are premature specifications as categories. That deciding that one is going to use them as the evaluative criteria makes certain presumptions about what is being evaluated which may not be correct assumptions either at that time or at any time.

The second point. That led us to a discussion of the realities of evaluative criteria, which we think are partially, first, that there are many different audiences, different users, multiple perspectives which each of these audiences have and therefore multiple and different criteria. That the values and criteria of these audiences must be specified in one way as one of the basic ways of defining evaluative criteria. That is one of the first jobs one has.

The third point. We then entered a discussion of the difference between criteria and measurement. If I can summarize that, we felt this a very important distinction because criteria are never subsumed by any single form of measurement. Criteria are the standards or objectives--that is, they are much more abstract than any single set of measures. The measures are the forms of information which are related back to the criteria.

Fourth point. We then turned from this relatively abstract discussion to a discussion of what are the concrete criteria that are now important in decision-making in various agencies--the agencies represented in the room. I am going to mention three of these.

One criterion which came up was compliance. That was the criterion being used by the Internal Revenue Service.

Another criterion which was discussed--and this one seemed to be extremely general across agencies--was influencing other organizations.

A third criterion was institutionalization of a program--that is, will someone else pick up the tab.

Let me reindicate what we are talking about here when we talk about criteria. These are the criteria that are used to determine what decisions will be made about. That is decisions are made on the basis of answers to these criteria. It is clear having presented these criteria to you, that most evaluators are not in the business of providing information relevant to them of doing measurement that is relevant to them. We thought that was quite interesting.

Someone in the group suggested that perhaps one other transcendent criterion that all agencies have is some aspect of cost-benefit analysis, or how much things cost.

Fifth, we then turned to a discussion of how to avoid what was characterized as Type 1 errors. That is, the consecutive shaping of the criteria in terms of what looked as though the measures would turn out to show to be successful.[40] This is the old problem of looking under the streetlight because the light is better there for the keys that you lost down the street where it was dark.

In other words, this is a danger raised about fitting one's criteria to one's successes and successively pruning along the way so that the measurement that was finally decided upon would be a very carefully selected set or single instance of gems rather than stones.

---

[40] See page 277, footnote 36, above.

The sixth and final point is that I think we all agree that evaluation should only be done where information will be used in decision-making, and that it was pointless to spend evaluation resources to conduct evaluations where the information would not be used in decision-making.

Trying to end on a hopeful note, we did that the following way. First, one of our group suggested that there are lots of simple questions which are essentially descriptive that can be answered and are being answered all the time. We are doing better at that.

Second, that a great many decisions require simple, descriptive information about what is happening and what is related to what. We are equipped to answer such things. It's only farther down the pike that we get to questions of why it is happening; and although we are very interested in those issues and those are the complex and policy-related issues, those are not the questions that are being asked.

We ended on a very opportunistic note when Kenneth R. Feinberg said that for the evaluator king who can come along and discover what programs will actually reduce the crime rate, the presidency is waiting. What I have given you is a list, and you might be interested in what we think is important and not important on that list. Perhaps I should leave that to the questions and to let other people in the groups--in the group that we had, answer.

MR. BENINGTON:

While people think of serious questions, I have a comment. I now see the statistician's view of the Constitution. And that is that the Executive makes Type 1 errors and the Legislative, Type 2.

MS. CHELIMSKY:

Are there any questions, serious or not?

PARTICIPANT:

I am Vickie Jaycox, the National Institute. My major complaint from this discussion goes back to what I feel is sort of a cop-out of evaluators at this point. Federal agencies, regardless of who the user of the evaluation is, at some point have to answer to whether or not they have had any effect on what the legislation was formulated for--basic questions of whether or not they are going to get refunded, related to whether their programs changed anything in the world. So when you get down to the question of whether Joe Shmo wants to refund his program, he has different questions to ask. When you are asking whether something changed in the world, then you are into a different kind of evaluation. Now, what everybody has been talking about is a very simple, straightforward, user-oriented evaluation. But there has been really no discussion of the role of really basic evaluation research, asking what is the effect of programs. I think it's something that was missed. I'd like some kind of comments on whether we are ever going to get back to real basic effectiveness evaluations, on whether we learned something conclusive from the evaluation. Does that make any sense?

MS. GUTTENTAG:

Absolutely. I hope I am free to give a personal opinion. Must I keep representing the panel?

Of course that is the question--does what we are doing matter in any way? What are the effects of what we are doing? I think that is the key issue in evaluation.

I personally though have been quite biased in reading evaluations by looking at the evaluation methodologies that have been used to answer that question. The methodologies themselves have often been inappropriate because of the assumptions that they have made, either about what is happening in the world, or about the statistical properties of what is happening in the world such that certain methods could be used. So, coming from that critical stance, I am always extremely concerned about what I call premature effectiveness evaluation. That is, it seems to me that more untruths have been told in the attempt to try to say what the effectiveness of a program is than the reverse. That is, I think we have been on safer ground in looking at a variety of different criteria and in keying evaluations to the criteria that decision-makers have so that the information that is produced is always in terms of the decisions that have to be made as a program develops. That is terribly abstract; I suppose that is why you get something of a bias in what I have said.

MS. JAYCOX:

I feel that it's because it's so difficult that we say, "Well, we don't want to do that anyway."

MS. GUTTENTAG:

My opinion is that we must provide decision-makers with information that they want.

MS. JAYCOX:

At a higher level, that is a very demanding kind of information. Did you reduce crime this year? We can't tell. We don't know.

MS. GUTTENTAG:

That's right. Well, you see, I think we are so much better off saying, We don't know, than, No, the things we did didn't help.

QUESTION FROM FLOOR:

Is it that you say,we don't know,or that you illuminate the number of things that affect the crime rate beyond the narrow things you measure?

MS. GUTTENTAG:

That is certainly one of the very useful ways of answering that question.

PARTICIPANT:

Walter Bergman, IRS. I share the same concern as that expressed by Ms. Jaycox. I think the answer is really long-term research as opposed to what I have only learned in very recent months or the last two years to know by the name of evaluation. Because it's a term we never even used. This takes more than answering a single administrator's politically motivated, generally immediate whim. I think it transcends administration. I think it transcends a single manager's interests. I don't think these answers can be gotten easily--I have to keep talking about IRS because it is something I know about. In our particular instance, we started in 1962 with our taxpayer compliance measurement program. We are trying to find out not what our body count is, but we are trying to find out whether we are doing anything to affect the public out there in terms of their behavior, their compliance. And what is happening to it. This does require a serious experimental design. We have had to develop panels. Unfortunately this means the same person gets audited twice in a row. We are trying to find out whether or not the fact that we audited him the first time made any difference in his behavior the second time. Fortunately, I would hope that our process is not considered destructive testing.

MS. GUTTENTAG:

Do you have informed consent to that?

QUESTION FROM FLOOR:

Informed by whom?

QUESTION FROM FLOOR:

Is it random assignment?

MR. BERGMAN:

Yes, it is random. It's random within random. We do get protests at times, I assure you. But somehow, we have been able to convince our constituency that this is necessary in order to maintain a voluntary compliance system with the tax system.

My only argument really is that I think we have to differentiate between long-run research, which will give us some insight into the real hard answers -- the final outputs that I mentioned before, versus some of the shorter-run evaluations. I think evaluations are wonderful for efficiency kinds of measurements. We do a lot of those, too.

MR. EWING:

Could I comment on that? I'd like to say that it seems to me that if you got at some juncture a willing ear on the part of a program manager or agency head -- whatever he might be -- decision-maker, somebody who is willing to talk about what his goals and objectives may be, one of the aspects of that situation is that you have got an interactive kind of discussion going, hopefully, in which he says what his objectives are; and you tell him what you can give him in the short run (if he wants something in the short run) and what you can't. You also tell him what can be measured currently

and what can't be or how well it can be. It seems to me that in keeping with our notion of a deal or a bargain, what you are hopefully able to work out is some kind of agreement that there are some things that can indeed be answered today or tomorrow or Friday or next week, and other things that will take a year and maybe will result in nothing much more than a disciplined judgment. Some other things require systematic analysis. Some things require monitoring only. Other things require evaluation of a fairly well-disciplined sort, and some other things can only be answered through long-term research.

Hopefully, you can work out therefore a kind of a mix of strategies which, combined, will begin not only to answer the perhaps politically motivated, short-term administrator's question, but also begin to serve the function of accumulating knowledge, putting building blocks in place and beginning to build a body of knowledge from which much more sensible judgments and decisions can be made.

To respond to an earlier question about where the agencies are that have done this, I don't know of any that have done it; but let me just say on behalf of one that has been much criticized, both here and elsewhere, that LEAA has at least put together an evaluation program which includes evaluation of its discretionary funds which are program evaluations, many of which are very clumsy and awkward kinds of evaluations. But it is also developing a program in the development of better methods for measuring and is also working on developing instructions to states and local governments on how to do simple evaluations and more complex ones and is also doing some long-range kinds of things including some things that started a couple of years ago -- for us it's long-range. For most Federal agencies it is. They are going to last another three or four years.

I think it is probably true that a great many agencies are working in that vein trying to put together a mix of strategies.

PARTICIPANT:

I'd like to make an observation, at least. You know, really, most of us who are here representing an agency are here because, to some varying degree, that agency is supporting an evaluation effort. In varying degrees, we are or have recognition within the agency. It seems to me one of the things we have on occasion in the conference overlooked is that we are ourselves in most cases managers who have evaluation as a product. We are responsible therefore as managers to really do a great many of the things that we are ourselves in turn talking about trying to get managers to do.

It seems on occasion we have to talk about negotiation. We have to negotiate our own products, all right. We have to sell those products as evaluation, and I think what we have had represented here also on occasion are a multitude of different management styles as well as evaluative styles. Some have reflected management styles that have worked. There are those, for instance, in education, who have said, you know, we have had an office that has been able to accomplish a certain sale of our product.

In other words, we are ourselves managers, and it seems we are mixing on occasion a management question with a product question. That has been a part of our conflict here.

PARTICIPANT:

I'd like to solicit your comments on how you feel about the same thing. I have gotten the feeling that there is sort of a projective need on the part of decision-makers, as compared to the

overall feeling of a retrospective view in evaluation. The decision-makers must take not only a retrospective view of what worked and what didn't work (and possibly why and how) but must also address the "what if" question. My decisions relate not so much to what has passed, but what is in the future: if I have options, or if I can identify alternative options, I need to have some means--sometimes it's the seat of the pants, it's intuitive, it's mathematical, call it what you will. All of those. But how do I convert what happened in that case, that set of cases, into the decisions I have to make about what will happen or what is likely to happen? How do I convert the retrospective into the projective?

We seem to have been focusing on what happened, and I don't know how we are going to get into crystal balls, map modeling, seat of the pants, how we are going to put these things together. But most of the decisions are not retrospective. They are projective. I'd like to get your feeling on how we convert an evaluation of a project that is on-going or that happened into projective tools that are credible?

MS. CHELIMSKY:

I think one of the big problems we have is that an evaluation's findings are often not generalizable (because of problems in the design, because of problems in the data, because of a million other reasons) even for the period in which they are derived. So that, you know, if they aren't even generalizable beyond the population studied to begin with, it's difficult to have confidence in their generalizability to unknown future situations.

PARTICIPANT:

As I see it, the basic problem of a manager is to generalize. And the point is, he either generalizes to different individuals, different programs, or to the same one continuing or changing. His is

inherently an objective task. I don't think we are facing the fact that our view has been essentially retrospective, but his view is essentially projective.

MR. EWING:

Let me comment briefly. It seems to me the usual scenario in most agencies is that people whose background is in research become evaluators or become the managers of evaluation which is contracted out. They get products which then get sent in nice neat packages to administrators of agencies, and the administrators don't read them because they are too thick or because they are untimely or because they simply have no training or background themselves which permits them to make head or tails of what is given them. Most administrators for some reason -- I'm sure there are reasons -- are not themselves trained in research or have any experience with research.

One of the things that is missing is a bridging function. We talked about that some in our panel. A function that involves somebody who understands enough about research to understand what it is the evaluation results say, but that same person has to understand enough about the needs of management to assure that he can take management's needs and make sense of them in terms of the evaluation results. That is a rare kind of person who can do that. It's a function that gets performed, I think, very seldom. It's one that I think most agencies have a great deal of trouble with, but it's not an impossible thing to do if somebody is assigned to do it who has some common sense. One of the troubles with it is that it hasn't been recognized well enough as a discrete function which needs to be performed and which is not typically well performed by a researcher or by a manager by himself.

Our administrator, for example, is fairly interested in evaluation results, but tends to be put off by them the more they are put in terms which he regards as research gobbledygook. That I think is a serious problem.

Related to that is a comment that was made in our panel which is that a great many people seem to make evaluation a very pretentious kind of thing. That is, more pretentious than it needs to be or deserves to be. If it were stated more modestly, it would not only be better understood, but more in keeping with the modesty of the findings. That might also help.

MS. GUTTENTAG:

There are of course models of inference which make it possible to take a prospective look. They are available.

III. WORKING PANEL I: IMPROVING THE INTERFACE BETWEEN AGENCY NEEDS AND EVALUATION

CHAIRMAN: CLIFFORD W. GRAVES,
Assistant Chief Administrative Officer and Director,
Office of Management and Budget
County of San Diego

MS. CHELIMSKY:

Cliff, what conclusions did Panel I come up with?

MR. GRAVES:

Well, the purpose of our panel was to suggest ways of improving communication between decision-makers and evaluators, on the assumption that such improvement would increase the use of evaluative information in decision-making.

The panel focused on the evaluator as the most adaptable party: the decision-maker and the decision-making process were taken as givens. Decision-makers make decisions and will continue to do so whether or not evaluation information is available. Furthermore, evaluation is only one of several kinds of information that decision-makers need: political, fiscal, legal and personal information are other kinds of appropriate input to the decision-making process. Evaluation supplements, but is not a sutstitute for, these other forms of evaluation. While this premise was not fully accepted by all panelists, we agreed that since evaluators appeared to perceive the problem more acutely, we would have to make the first move.

The panel approached its task as a market research problem. Evaluators have the capacity to supply a product (or service). This capacity is not infinite, but it is ample:

- The state-of-the-art is highly advanced.
- The Federal government spends substantial funds for evaluation.
- There is a large supply of trained personnel within and available to the government.

(Again, this premise was not fully accepted by all panel members but the disagreement was only a matter of degree.)

Market interest (potential demand) appears to be present. Increasingly, decision-makers talk as though they would like to have evaluation information; they have supplied increasing resources and status for the evaluation function. However, decision-makers may not fully understand what evaluation is and what it can do.

Decision-maker interest in evaluation is more the result of the growing complexity and openness of the decision-making process and a growing awareness of the shortcomings of other forms of information. In short, decision-makers are interested in evaluation not because they understand what it is, but rather because of the changing environment in which decisions are made.

The panel also accepted the premise that evaluation information will not be used unless the decision-maker wants to use it and that evaluation information should not be used unless the decision-maker knows how to use it.

So, given our ability to supply, and evidence of a market for our product, how should we proceed?

THE APPROACH: FIVE ISSUES

The panel addressed five issues in its search for a "marketing strategy."

1. Who are the decision-makers? Where do they come from and what kind of decision-making environments exist?

   The use of the term "decision-maker" tends to obscure the fact that there are many kinds of decision-makers operating at many points in the process. They vary in the kinds of decisions they can (or are willing to) make, their perspective on any given issue and, therefore, the types of evaluation information they may require. We looked at this issue in two ways.

   First, we focused on the concept of the environment of an agency (or government as a whole) as a decision-making system. The panelists believe that decision-makers are part of a larger system and it is this system, rather than the individual decision-maker, that must be understood. For example, the Congressional Budget Act establishes a system of related decisions and assigns responsibility for those decisions among various elements of the Congress and the Executive Branch. By understanding that Act and the decisions it requires, the key points where evaluation information can be helpful can be readily identified. Similarly, there is a system within each Executive Branch agency.

   The first step in designing an evaluation strategy should be to understand the organization and functioning of the system of interest. Once that is understood, evaluation systems should be designed to fit.

We also discussed individual decision-makers, who come in all shapes and sizes. We were intrigued by a suggested distinction made the first day of this symposium, between "decision-makers" and "position-takers."[41] The latter are the persons who absorb and analyze information and then package it for decision-makers. Position-takers are found in large numbers in the Federal government, and constitute a good market for evaluation information. They have more time than decision-makers, and a better understanding of the analytical side of the evaluation process. They are conduits for the flow of evaluation information to decision-makers.

The panel touched on the cascade characteristic of governmental decision-making. At the top are the legislative and high-level policy-making processes involving relatively few people and very coarse-grained decisions. These decisions in turn cascade down through the organization, setting off administrative decisions. At each level of the cascade, there is potential demand for evaluation information; different types are needed, ranging from broad impact and inter-program effectiveness issues at the upper levels down to operational efficiency-type questions toward the bottom.

The panelists briefly discussed the importance of understanding the incentives that guide the actions of decision-makers. Much has been written concerning the short-run outlook of decision-makers, usually in an oversimplified way. However, it is important when addressing the evaluation

[41] See pages 101-103 above.

needs of particular decision-makers or decision-making levels, that the evaluator understand what makes a decision-maker tick.

Within the panel, a minority view took issue with the panel's approach. That view pictured the evaluator as the seeker of truth, independent of the decision framework, letting chips and decisions fall where they may.

2. What distinguishes evaluation information from other kinds of information needed and/or used by decision-makers?

Evaluation information is neither better nor worse than other types of information; it is simply different. The panel spent quite a bit of time trying to determine just what distinguishes evaluation information from other types, and finally identified the following characteristics:

- It is structured information, set within a context, clearly circumscribed. This creates problems of distilling evaluation findings into executive summaries, news articles, and the like; because the first thing to go in such distillations is the context.
- It sets confidence limits, by including cautions to the users.
- It describes and answers questions about a real activity or set of activities according to some theory. It is retrospective, and it addresses specific questions.
- It describes effects against some standards. In fact, often the most important contribution of evaluation projects is the establishment of standards against which activities can be measured.

- It is a feedback loop in a continuing process of program development, execution, and refinement. In this sense, it is use-oriented.
- Its scope includes consideration of the side effects of a given activity, i.e., it is not a closed-end analysis.
- It does not assign values to a given activity, but rather tests the activity against values assigned by others.

The panel concluded that the methodology of evaluation is not its distinguishing characteristic. Evaluation makes use of many techniques common to other forms of research and analysis.

3. Evaluation's potential contribution is not fully comprehended by decision-makers; is this because its products are badly designed, badly packaged, directed at the wrong segment of the market, poorly advertised, or something else?

As used here, "badly designed" means directed at questions of little interest to the decision-maker, or otherwise structured to yield irrelevant information. "Badly packaged" means that evaluation is not presented in a usable or recognizable form. "Wrong segment of the market" means that the evaluation is not directed at decision-makers or is directed at the wrong decision-makers. "Poorly advertised" means that the decision-makers are not aware of the existence of the information or are unaware of its potential value.

We concluded that the answer was "yes" to all of these. We then went on to focus on the notion of evaluation as a threat. The panel believes that the threatening nature of

evaluation may be the most important obstacle to its effective use. Evaluation is a threat because:

- It is public information which, once generated, cannot be kept secret or limited to the private use of a decision-maker. Thus, it provides persons other than the responsible decision-maker with information which may adversely affect that decision-maker.
- It is a change force: it seeks ways to improve (change) an existing set of activities. Change is inherently threatening.

To overcome this, the panel believes in the importance of including "victims" in all phases of evaluation projects from pre-design and planning through execution and product packaging. The theory is, the more a project is seen to be controlled, or primarily usable by the program manager, the more likely the evaluation information is to be accepted when it is completed. Also, the more a program manager or decision-maker knows about how an evaluation project was put together, the better he is able to implement the changes recommended. Not incidentally, by giving the program manager a head start before making evaluation information public, he is able to accept and perhaps claim credit for identifying ways to improve his program. As an example, the Environmental Protection Administration does this through a task force approach to most evaluation projects.

In the opinion of many of the panel members, evaluation as practiced at the Federal level is now 99% production and 1% marketing. The lack of attention to marketing (who needs it and how can it be used?) is a major shortcoming.

While apparently common sense, market research is rarely done before a project is started. This may be because of poor communication between the evaluation and the decision-making processes and some uncertainty at the evaluator's level about intended uses.[42] Even the timing (the point at which evaluation information may be necessary) is not always clear. These are obstacles to be overcome, however, not excuses.

The panel also agreed that the higher up you go in the decision-making hierarchy, the less time the evaluator has to present evaluation information and the less the stability of the decision-making environment. This has two implications. First, the evaluator should aim at the more stable elements of the decision-making process, such as Congressional staff (position-takers) and program managers, rather than an individual Senator or a Cabinet officer. The second implication is that in order to secure a significant amount of the time of the top-level decision-makers, you must get their attention. This can only be done if they are aware that a real problem exists. The experience of panel members was that if the top-level decision-maker is aware that there is a problem, that decision-maker will take whatever time is necessary to review information that might lead to a solution.

---

[42] See Issue 1, pages 358 through 360, also 334 through 336 above.

4. What are the criteria for measuring the effectiveness of evaluation products: technical quality, timeliness, acceptance of recommendations, state-of-the-art advancement, or others?

We've passed the point, as evaluators, where we believe the only standard of a "good" evaluation is whether the recommendations coincide with an actual decision. However, if this is not the principal standard, what other criteria should be used? After all, an evaluation program merits evaluation just as other programs do.

The panel came up with an interesting notion: an evaluation can be considered a success if, according to the evaluator's measures, the program evaluated subsequently improved. The idea here is that, in analyzing the subject program, the evaluator identified or clarified appropriate measures for program performance. If, following the evaluation, the subject program performance improved according to those measures, then the evaluation can be considered a success. (If the evaluation of the program showed that the program was already a total success, then a continued high level of performance against those measures would be acceptable).

To make this assessment requires follow-up to an initial evaluation project, and continuing involvement of the evaluator in the program. This is not usually the case in Federal evaluation programs.

This notion flows out of one of the characteristics of evaluation information noted earlier, that is, it is part of the continuing process of program development including planning, implementation and evaluation.

Other factors we identified as important criteria were timeliness and availability (being in the right place at the right time) and whether they raise the consciousness of persons associated with the program to issues of program performance.

5. Given answers to the above, what are the responsibilities of the evaluator, the evaluation manager, the research community and the decision-maker in developing an acceptable product?

There are many players in the evaluation game, each of whom bears some responsibility for an acceptable evaluation product. We kicked around the idea of mandated processes such as the Congressional Budget Act, OMB Circulars, and some of the pending sunset legislation. These have the initial attraction of being action-forcers. However, the panel was not enthusiastic about these as forces to improve the quality and usefulness of evaluation products. Mandated processes guarantee large quantities of evaluation production, but not high quality.

The panel also concluded that evaluators can't do it all, although they can stimulate improvement. The evaluator accepts and conducts assignments but has no institutional responsibility for evaluation planning or for the utilization of evaluation products.

We zeroed in on the evaluation manager--the person or unit responsible for planning, packaging, and disseminating evaluation findings--as the critical factor to the development of an acceptable product. The evaluation manager is the bridge between evaluator and decision-maker. This is the point from which most of the marketing needs to be done; this is the interface.

The panel also noted that the role of the evaluator changes depending on the skills in the agency (assuming that most evaluation is done by outside contractors). Some agencies have a highly sophisticated evaluation process, in which cases the evaluator is more the arms and legs of the agency, carrying out projects designed and pre-marketed within the agency. On the other hand, other agencies lack this sophistication and are, in effect, buying brains, as well as arms and legs. One panel member lamented that the cost per unit of evaluation information should be higher in the latter case, but Federal contracting processes do not recognize the difference.

One of the panel members developed a specific assignment of responsibilities for each of the players in the evaluation game which the panel concluded was a very good one. That report[43] follows directly after this.

CONCLUSIONS

Out of its deliberations, the panel was able to distill its concerns down to just a few points.

The first and most important one is that the approach to evaluation in each agency or decision system must fit that agency or system. There are no universal truths to the design and conduct of evaluation and no universal characteristics to the market. In other words, each evaluation program must be tailor-made.

[43] See pages 371 - 373 below.

Second, the most important step toward improving the interface is understanding it. This analysis of the environment and the potential contribution of evaluation is absolutely critical.

Third, evaluation should be seen as part of a continuing loop of program operations, including planning and implementation and evaluation. It is not a separate outside force. Unfortunately, this is not recognized in most agencies.

Having come to these apparently common-sense truths, the panel then concluded that there was little to be gained by further exploring any of them as generalities. Nevertheless, these represent a major agenda for all persons concerned with the effective utilization of evaluation to improve the quality of Federal decision-making.

MS. CHELIMSKY:

Do we have some questions? Comments?

PARTICIPANT:

I am Joel Garner, LEAA. I would like as an evaluation manager, or at least as project monitor for evaluation, to reject the idea that I am responsible for bringing coherence to the chaos that we find in terms of the relationship between evaluators and programs. If my office or my personal advancement in the agency is based on that kind of assessment, I need to put out more resumes. I would also like to reject the idea that evaluation is to be assessed on whether the program we are evaluating is in fact successful. Again, if my office or my personal advancement is based on the ability of LEAA to reduce the crime rate (I believe you said that the program itself has to

improve after the evaluation), then, if the program doesn't reduce crime more after the evaluation, the evaluation was not successful. That's the way I read it. If that's what you said, I think that's a very dangerous thing to say.

MR. GRAVES:

Let me clarify, but before I do, I can't resist going after your first assertion. If improving programs through evaluation is not your responsibility, what the hell is your responsibility?

MR. GARNER:

Well, it's not solely the evaluation manager's responsibility. There are other people who can be blamed.

MR. GRAVES:

Going onto the second point, what I was trying to get across (as far as the determination of what an effective evaluation may be is concerned) is this. If you include utilization as somehow part of your evaluation criteria, one of the products of evaluation is a set of measures, perhaps a refinement of measures which already existed. This is a way of looking at whatever program you are evaluating. That is really the first thing you do in an evaluation project, and then you proceed to measure the program's performance against those standards. You come to certain conclusions about it.

Our view was that if you have done that and you have an effective evaluation, then the program that you evaluated should somehow perform better against those measures after you did the evaluation.

Joe Nay sparked this notion with our panel. Perhaps he'd like to amplify these comments.

MR. NAY:

Yes, in answer to crime, I'd say, well, you shouldn't pick crime. It's not a good measure. I'd say that, having agreed on a legitimate set of measures, then you ought to be able to expect to see improvement in subsequent periods if the program goes on, in those measures that were used. If crime is a bad measure--and it is for many programs--then that shouldn't be the measure you are using. You get that at the beginning--wipe that out at the beginning, not at the end.

MS. CHELIMSKY:

It seems to me that Blair's point was that there may be more actors with more roles than Cliff and his panel have suggested. That is, they are suggesting there is an evaluator, an evaluation manager and a decision-maker; and there may be a whole lot of other people who are critical to that process including perhaps somebody between the evaluation manager and the decision-maker whom we talked about as being a kind of bridge-builder, interpreter, translator--whatever you want to call him or her.

There are also the people who plan the program and the people who receive the results of the evaluation, both of whom have some responsibility for seeing to it that the things that are designed are things that can be evaluated, at least to some degree. They have that responsibility. There are people who then have to take those results and make use of them. So I would suggest that there may be two or three more players in the game.

MR. GRAVES:

I'd agree with Blair. But the point we are trying to make is that, as a matter of fact, I gathered from looking at the roster of people here that most of the people here are evaluation managers or in-house evaluators--somehow responsible for evaluation in-house. There is always a tendency to blame somebody for a problem who isn't around. We had a question early in our panel in terms of why did we have to accept the decision-maker as a given. It's an "all-his-fault" kind of thing.

I made the comment that it's easy to blame the decision-maker because he's not here. But, in my opinion as an ex-Fed, the role and mission of the evaluation division, the Assistant Secretary for Evaluation--whatever it happens to be in an agency--is never clear. I think most of us ended on an optimistic note that maybe one of the things that would come out of this symposium--at least some of the ideas we had--was a clarification, a better understanding on the part of the evaluation manager in terms of what his role and responsibility is.

QUESTION FROM FLOOR:

I would like to back up to Joe's second question. I think that we have taken the assumption that all evaluation is critical, condemnatory. Once in a while on a rare occasion, we find research that is not critical, but that is, in fact, supportive and does not lead to the kind of feedback you are talking about in that we are supporting a homeostatic situation.

MR. GRAVES:

In that case I would say that the criterion would be that it not get any worse, as a result, after the evaluation!

ATTACHMENT TO THE REPORT
OF WORKING PANEL I

ASSIGNMENT OF RESPONSIBILITIES IN THE GAME OF EVALUATION

JOE N. NAY, The Urban Institute
(Member of Working Panel I)

PLAYER: DECISION-MAKER — RESPONSIBILITIES:

- DEVELOP (WITH HIS EVALUATOR) AN UNDERSTANDING OF THE DECISION-MAKER'S OWN ROLE, NEEDS, AND MEASURES.
  - ACCEPTABLE MEASURES
  - BELIEVED PROGRAM LOGIC
  - MANAGER'S ABILITY, AUTHORITY, INTENTIONS TO ACT.
- PARTICIPATE IN CYCLIC CLOSING OF GAPS BETWEEN BELIEVED PROGRAM LOGIC AND ACTUAL PROGRAM LOGIC.
- TAKE TIME TO UNDERSTAND THE RESULTS. (MANAGERS SHOULD NOT EVALUATE THINGS WHOSE RESULTS THEY WON'T TAKE TIME TO STUDY).
- PARTICIPATE IN DECISIONS ON SEQUENTIAL PURCHASE OF INFORMATION.

PLAYER: EVALUATOR

- UNDERSTAND BOTH THE RHETORICAL PROGRAM AND THE ACTUAL PROGRAM.
- TIGHTLY RELATE ISSUES, MEASUREMENT POINTS AND MEASURES, COMPARISONS, AND STRUCTURE OF THE MANAGEMENT AND INTERVENTION PROCESS.
- DO ENOUGH PARTICIPANT OBSERVATION TO KNOW WHAT IS REALLY HAPPENING.

ASSIGNMENT OF RESPONSIBILITIES IN THE GAME OF EVALUATION
(CONTINUED)

PLAYER: EVALUATOR (CONTINUED)

RESPONSIBILITIES:

- PROVIDE ENOUGH STRUCTURE (FLOW DIAGRAMS?) TO SHOW HOW THE MEASUREMENTS TAKEN ARE INTERRELATED THROUGH THE ACTUAL PROCESS ACTIVITIES.
- CAPTURE EXOGENOUS VARIABLES AND INTERNAL FEEDBACK LOOPS.
- BE HEAVILY INVOLVED IN ACTUAL MEASUREMENT.
- CHOOSE APPROPRIATE ANALYTIC TECHNIQUES FOR PRODUCING INFORMATION FROM DATA.
- MEASURE, MAKE COMPARISONS, PRODUCE INFORMATION, EXPLAIN IT. MAKE RESULTS ACCESSIBLE TO VARIOUS LEVELS.

PLAYER: EVALUATION MANAGER (IN THE AGENCY)

- HAVE AT LEAST A FEW PEOPLE WHO ARE COMPETENT TO DO THE WORK THEMSELVES.
- INVOLVE THEMSELVES IN THE ENTIRE LOOP OF MANAGEMENT, INTERVENTION, AND EVALUATION SO THAT THEY ARE KNOWLEDGEABLE IN ALL PARTS OF IT.
- FACILITATE AND REQUIRE INTERFACES AT MANAGER/EVALUATOR AND DIRECT INTERVENTION/EVALUATOR LEVELS.
- DO MARKET ANALYSIS AND ASSESS POTENTIAL USERS AND USES, POLICY MARKET, PROGRAM MARKET, INDIVIDUAL MARKET.
- DON'T BE AFRAID TO STRUCTURE THE WORK THAT YOU WANT, GET PEOPLE WHO CAN DO IT, AND REQUIRE THEM TO.

ASSIGNMENT OF RESPONSIBILITIES IN THE GAME OF EVALUATION
(CONCLUDED)

PLAYER: EVALUATION MANAGER (CONTINUED)

RESPONSIBILITIES:

- DO:
  - EVALUABILITY ASSESSMENTS,
  - ISSUES ANALYSIS,
  - FIELD WORK,
  - SYNTHESIS OF TESTABLE RHETORICAL AND OPERATING MEASUREMENT MODELS,
  - ASSESSMENT OF WHAT IS KNOWN, AND
  - DESIGNS AND COSTS FOR KNOWING MORE.
- BE THE AGENCY'S CUSTODIAN OF A CONTINUING STORE OF MODELS, KNOWLEDGE, RESOURCES.

CONCLUDING REMARKS

ELEANOR CHELIMSKY,
Department Head, Program Evaluation,
METREK Division of The MITRE Corporation

MS. CHELIMSKY:

Time is fleeting; we're all tired. Let me sum up quickly. I guess in trying to crystallize what I feel has been said at this symposium about evaluation over the last three days, I keep thinking about the citizen reactions to police response time that George Kelling was talking about yesterday. The idea was that if you expect the police to come in five minutes and they get there in ten, you're disappointed. But if you expect them in ten and they get there in five, you're elated.

It may be that much of our dissatisfaction with evaluation today lies not in evaluation, but in ourselves. Many of us find that we are burdened with transactions and activities, that we have less and less time or energy or talent or inclination left at the end of the day either for reflection or for communicating our thoughts adequately. We may be counting on evaluation to fill gaps it was never intended to fill. I think if we expect evaluation to be a surrogate for thinking, as Jim Gregg said, or for problem solving, or for communicating with others, we are going to be disappointed.

If, on the other hand (as Dan Wilner said to me last night in the corridor), we contrast where we are today--in terms of getting acceptance for uses of evaluation--with where we were ten years ago, there is some cause for elation.

It has been said many, many times over the past three days that evaluation is only a tool, but it does allow something infinitely precious--the bringing of some rationality into areas which are heavily charged and counter-charged with emotion and with self-interest. Clearly we are not going to dissipate all those war-clouds with one small ray of evaluative enlightenment, nor should we expect to. We need, as John Evans and Joe Lewis have said, to accumulate evidence patiently and to help it develop its own momentum.

I think there has been some fruitful airing of divergent, long-term goals and aspirations for evaluation among us. There has been perhaps less airing of how to get there from here. We heard Tom Kelly give some useful definitions of decision-making and position taking, yet threatened program managers remain a major obstacle for the integrity of evaluation, for the accumulation of evidence.

We received clear statements from OMB and Congress about their uncompromising intentions to aggressively pursue the use of evaluation in order to strengthen their review and oversight functions (and of course, I am thinking here of Joe Nay's first definition of the term, not his second). What is less clear, however, is how they intend to do all that; that is, what incentives and sanctions can be, or will be, used in this area?

The goal of this conference was to confront various points of view about where we are today in evaluation and to confront them with candor. I think we have done this, but I am not sure we have done it completely. Some of the workshops were too big, perhaps, or another format was needed.

On the other hand, evaluators have not been shy about reproaching agencies with their managerial sins, and with other sins as well, both of omission and commission. Agency people have accused evaluators of

gross misdemeanors such as irrelevance, untimeliness, triviality, jargon and over-theologizing, as well as of leading innocent administrators like Chuck Work down the garden path. But evaluators and agency people have also blamed themselves for their own failures. I think the self-deprecating note struck by Jerry Caplan when he articulated the theme of this conference was very helpful. He set the tone for whatever honesty and humility we may have been able to achieve here.

In closing, I'd like to express my appreciation for what I found to be a very open and intellectually stimulating set of statements and interventions by the people here. I know that any conference is the sum of its participants; and if this one has been interesting, it's unquestionably because of the people who have been kind enough to lend us their presence here. Thank you all very much. Cultivate your garden, as Voltaire said, and Bon Voyage to all of you.

END

gross misdemeanors such as irrelevance, untimeliness, triviality, jargon and over-theologizing, as well as of leading innocent administrators like Chuck Work down the garden path. But evaluators and agency people have also blamed themselves for their own failures. I think the self-deprecating note struck by Jerry Caplan when he articulated the theme of this conference was very helpful. He set the tone for whatever honesty and humility we may have been able to achieve here.

In closing, I'd like to express my appreciation for what I found to be a very open and intellectually stimulating set of statements and interventions by the people here. I know that any conference is the sum of its participants; and if this one has been interesting, it's unquestionably because of the people who have been kind enough to lend us their presence here. Thank you all very much. Cultivate your garden, as Voltaire said, and Bon Voyage to all of you.